

Mathematics of AI safety

Xiaolong Han

DEPARTMENT OF MATHEMATICS, CALIFORNIA STATE UNIVERSITY, NORTHRIDGE, CA
91330, USA

Email address: `xiaolong.han@csun.edu`

ABSTRACT. AI safety is the study of how to make artificial intelligence (AI) beneficial and how to prevent the harms that it may cause. The first two of its main concerns are the misuse of AI systems by people and misalignment: an AI system may pursue a goal different from the intended one, either because the goal was specified incorrectly or because training instilled a different goal. The others are a lack of robustness, i.e., failures on inputs unlike the training data; a lack of legibility and steerability, i.e., humans can neither understand how a system computes nor reliably correct it; failures of cooperation among AI systems and humans; and the gradual disempowerment of humans, who delegate more and more decisions to systems that they do not understand.

In these notes, we focus on one aspect, which lies between misalignment, robustness, and legibility: among the many parameters that fit the training data, which one does training select, and when? We formulate this question in mathematics, and we study the problems that it raises in singular learning theory, in the dynamics of training, and in the metastability of diffusions, where it becomes a semiclassical problem for the Witten Laplacian. We also study superposition, i.e., how a network represents more features than it has dimensions, and the Fourier analysis of the algorithms that networks learn for modular addition; both connect legibility with harmonic analysis.

Preface

In these notes, we introduce a family of problems in mathematics that arise from the safety of artificial intelligence (AI). The notes are divided into three parts.

In Part I, we explain what modern AI is in mathematics language: data, models with parameters, losses, neural networks, and training by gradient descent. We use a small neural network that recognizes handwritten digits as a running example, and we describe the language models on which chat systems are built. We end with the deep linear network, the one model in which training can be solved explicitly.

In Part II, we explain what AI safety is, and why the question “which solution does training find, and when?” is a safety question.

In Part III, we develop the mathematics needed to study the problems: singular learning theory, the staircases of training, diffusions and metastability, the Eyring-Kramers law for singular wells, entropic selection and the noise of training, superposition, and the Fourier analysis of learned algorithms. Among the many open problems on mathematics for AI safety, we chose those which are closest to semiclassical analysis, spectral theory, and harmonic analysis.

We fix the following notation throughout the notes.

- d is the number of parameters, $w \in W \subset \mathbb{R}^d$ (or $w \in \mathbb{T}^d = \mathbb{R}^d / (2\pi\mathbb{Z})^d$) is a parameter, and W_0 is the set of optimal parameters.
- x is an input, y is an output, and n is the number of data points (the sample size). Random data points are written $X_1, \dots, X_n$.
- f_w is the function computed by a model with parameter w , and σ is an activation function.
- L is the loss, normalized so that $L \geq 0$ and $L^{-1}(0) = W_0$; L_n is the empirical loss computed from n data points.
- q_0 is the true distribution of the data, $p(\cdot | w)$ is the model, φ is a prior density, and F_n is the free energy.
- λ and m are the learning coefficient and its multiplicity.
- t is time, η is the learning rate, ε is the temperature, and w_t is a (random) training trajectory.
- H is the width of a hidden layer, A and B are weight matrices, Φ is a target matrix with singular values $s_1 > s_2 > \dots > s_r > 0$, and a_α is the strength of the α -th mode.
- π_ε is the Gibbs measure, τ_C is the hitting time of a set C , $q_{A,B}$ is an equilibrium potential, cap_ε is a capacity, and h is a barrier height.
- $\mathbb{E}$ and $\mathbb{P}$ denote expectation and probability; $\mathbb{E}_w$ and $\mathbb{P}_w$ refer to a process started at w .
- j, k, l are indices.

Contents

Preface	i
Part 1. Artificial intelligence	1
Chapter 1. Learning from data	2
1.1. Supervised learning	2
1.2. Statistical models and likelihood	3
1.3. Entropy and the Kullback-Leibler divergence	6
1.4. Fisher information and singular models	7
1.5. The Bayesian point of view	9
1.6. Generalization	12
Chapter 2. Neural networks	14
2.1. Layers and neurons	14
2.2. Symmetries and singularities	16
2.3. Language models and agents	17
Chapter 3. Training	18
3.1. Gradient descent and gradient flow	18
3.2. Stochastic gradient descent	19
3.3. The Langevin dynamics as a model of noisy training	21
3.4. Plateaus and stages	22
3.5. The deep linear network	23
Part 2. AI safety	27
Chapter 4. What is AI safety?	28
4.1. Risks from powerful AI	28
4.2. Where things go wrong	28
4.3. Interpretability	29
4.4. Generalization in and out of distribution	30
Chapter 5. Which solution does training find?	31
5.1. Many solutions fit the data	31
5.2. Two accounts of selection	32
5.3. Stages of learning and developmental interpretability	33
5.4. Two clocks	34
5.5. From questions to problems	35
Part 3. Mathematics of AI safety	36
Chapter 6. Singular learning theory	37

6.1.	Bayesian learning and the free energy	37
6.2.	The regular case	38
6.3.	The learning coefficient	39
6.4.	The local learning coefficient	46
6.5.	The stratification by the learning coefficient	50
6.6.	Computing learning coefficients	51
Chapter 7.	Staircases	54
7.1.	The Bayesian ladder	54
7.2.	Opposing staircases	55
7.3.	The free energy of a saddle	56
7.4.	Effective losses	59
Chapter 8.	Diffusions and metastability	62
8.1.	Brownian motion and the Langevin dynamics	62
8.2.	Hitting times and boundary value problems	65
8.3.	Metastability in dimension one	66
8.4.	The Eyring-Kramers formula	68
8.5.	Capacities	71
8.6.	The semiclassical point of view	72
Chapter 9.	Eyring-Kramers laws for singular wells	75
9.1.	The problems	75
9.2.	The Gibbs mass of a singular well	76
9.3.	Capacity through a Morse gate	77
9.4.	An averaged singular Eyring-Kramers law	78
9.5.	The pointwise problem	80
Chapter 10.	Entropic selection and the noise of training	82
10.1.	Two strata of equal loss	82
10.2.	Degenerate noise	85
Chapter 11.	Superposition	87
11.1.	The hypothesis	87
11.2.	Room	88
11.3.	The price	89
11.4.	Decoding	91
11.5.	Beyond the square-root bottleneck	92
11.6.	The toy model	97
11.7.	Problems	99
Chapter 12.	Fourier analysis of learned algorithms	104
12.1.	Fourier analysis on $\mathbb{Z}/p\mathbb{Z}$	104
12.2.	Exact fits by single-frequency neurons	105
12.3.	The minimal width	107
12.4.	Learning coefficients and the Fourier mechanism	110
12.5.	Which frequencies are learned	113
Chapter 13.	Where to start	115
13.1.	Metastability of singular wells	115
13.2.	Laplace integrals in singular learning theory	115

13.3. Harmonic analysis of superposition	116
13.4. Fourier analysis on $\mathbb{Z}/p\mathbb{Z}$	116
13.5. A reading course	117
Appendix A. Glossary	118
Appendix. Bibliography	121

Part 1

Artificial intelligence

CHAPTER 1

Learning from data

Artificial intelligence (AI) refers to computer systems that perform tasks which we associate with intelligence, such as recognizing images, translating languages, writing programs, and proving mathematics theorems. Modern AI is dominated by *machine learning*. Instead of writing down the rules of a task by hand, one specifies a large family of candidate programs, indexed by real parameters, and selects the parameters from examples of the task. Every AI system discussed in these notes is produced in this way.

In this chapter, we describe this framework in the language of mathematics: data, models, losses, statistical models, entropy, Fisher information, the Bayesian point of view, and generalization. Several of these notions come from statistics, information theory, and statistical mechanics, and we explain their original meaning before we explain how they are used in machine learning. We use a small neural network that recognizes handwritten digits as a running example.

1.1. Supervised learning

The simplest form of machine learning is supervised learning: we are given examples of inputs together with the correct outputs, and we want a program that produces the correct output for a new input.

Example (Handwritten digits). A grayscale image of $28 \times 28 = 784$ pixels is encoded as the vector $x \in [0, 1]^{784}$ of its pixel intensities, where 0 is black and 1 is white. The task is to recognize the digit $y \in \{0, 1, \dots, 9\}$ written in the image. A standard data set is the MNIST (Modified National Institute of Standards and Technology) database of 70,000 images of handwritten digits, labeled by humans; see LeCun-Bottou-Bengio-Haffnerⁱ. Of these images, 60,000 are used for training and 10,000 for testing.

Definition (Data). Let $\mathcal{X}$ be a set of inputs and $\mathcal{Y}$ be a set of outputs. A data distribution is a probability measure on $\mathcal{X} \times \mathcal{Y}$. A training set of size n is a finite sequence

$$\mathcal{D}_n = \{(x_j, y_j)\}_{j=1}^n \subset \mathcal{X} \times \mathcal{Y}$$

of independent samples from the data distribution.

The data distribution describes how often each input occurs in the world, together with its correct output. It is never known; only the samples in $\mathcal{D}_n$ are.

Example (Handwritten digits). Here, $\mathcal{X} = [0, 1]^{784}$, $\mathcal{Y} = \{0, 1, \dots, 9\}$, and the data distribution is the distribution of the pairs (image, digit) over all digits that people write. The 60,000 training images of the MNIST database form a training set of size $n = 60,000$.

Definition (Models and parameters). A model is a family of functions $f_w : \mathcal{X} \rightarrow \mathbb{R}^N$ indexed by a parameter $w \in W$, where the parameter space W is a subset of $\mathbb{R}^d$. The number d is the number of parameters of the model.

ⁱY. LeCun, L. Bottou, Y. Bengio, and P. Haffner, *Gradient-based learning applied to document recognition*. Proc. IEEE 86 (1998), 2278–2324.

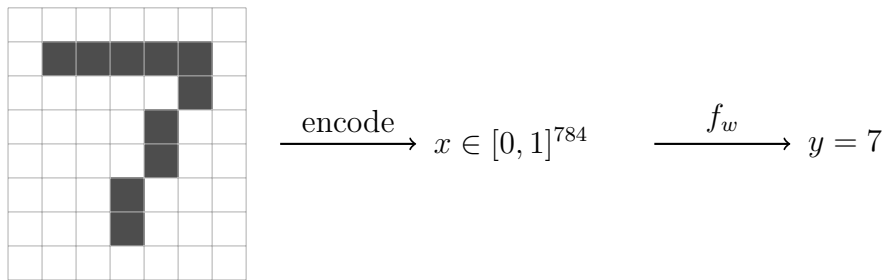


FIGURE 1.1. A handwritten digit, encoded as a vector of pixel intensities (schematically, with fewer pixels), and its label

The output $f_w(x)$ is a prediction: either the predicted output itself, or a vector of scores from which a prediction is read off.

Definition (Losses). A loss function is a function $\ell : \mathbb{R}^N \times \mathcal{Y} \rightarrow [0, \infty)$, where $\ell(f_w(x), y)$ measures the error of the prediction $f_w(x)$ when the correct output is y . Given a training set $\mathcal{D}_n$, the empirical loss is

$$L_n(w) = \frac{1}{n} \sum_{j=1}^n \ell(f_w(x_j), y_j),$$

and the population loss is

$$\bar{L}(w) = \mathbb{E}[\ell(f_w(x), y)],$$

where (x, y) is distributed according to the data distribution.

Remark (Empirical and population losses). The empirical loss $L_n(w)$ is computed on the training data only: it is the average error of f_w over the n training examples. It can be evaluated, since the training data are at hand.

The population loss $\bar{L}(w)$ is the average error of f_w over all possible inputs and outputs, weighted by the data distribution. It measures the performance of f_w on data that have not been seen. It cannot be evaluated, since the data distribution is unknown.

For each fixed w , the terms $\ell(f_w(x_j), y_j)$ are independent and identically distributed with mean $\bar{L}(w)$. So by the strong law of large numbers,

$$L_n(w) \rightarrow \bar{L}(w) \quad \text{almost surely as } n \rightarrow \infty.$$

However, training chooses w depending on the training data, and for the chosen w , the empirical loss may be much smaller than the population loss. This gap is the subject of Section 1.6.

For $\mathcal{Y} \subset \mathbb{R}^N$, the *square loss* is

$$\ell(y', y) = |y' - y|^2.$$

Training a model means choosing w with small empirical loss $L_n(w)$, usually by the iterative algorithms of Chapter 3. The actual goal is a small population loss $\bar{L}(w)$.

1.2. Statistical models and likelihood

So far, a model produces a single prediction for each input. It is often more useful to let the model express its uncertainty, by a probability distribution over the possible outputs. In Sections 1.2–1.5, we describe this point of view. Throughout, a probability density on $\mathcal{Y}$ is taken with respect to the counting measure if $\mathcal{Y}$ is finite, and with respect to the Lebesgue measure if $\mathcal{Y} = \mathbb{R}^N$. In the first case, a probability density is a probability vector, and integrals over $\mathcal{Y}$ are sums.

A model with N scores $f_w(x) = (f_1, \dots, f_N)$ is turned into a probability distribution on N outputs by the softmax:

$$p(k \mid x, w) = \frac{e^{f_k}}{\sum_{l=1}^N e^{f_l}} \quad \text{for } k = 1, \dots, N.$$

These numbers are positive and sum to 1, and a larger score gives a larger probability. The scores f_k are called *logits*.

Remark (Gibbs distributions). The softmax is a Gibbsⁱ distribution of statistical mechanics. Consider a physical system with finitely many states k of energies E_k . In equilibrium at the temperature $T > 0$, the system is in the state k with probability

$$\frac{e^{-\frac{E_k}{T}}}{\sum_l e^{-\frac{E_l}{T}}}.$$

Low energies are likely, and the temperature measures how much the system fluctuates. The softmax is the Gibbs distribution with the energies $E_k = -f_k$ at the temperature $T = 1$. As $T \rightarrow 0$, the Gibbs distribution concentrates on the states of lowest energy, i.e., on the largest scores. This is why the softmax is a “soft” maximum. The same Gibbs distributions, with the loss as the energy, are the stationary laws of noisy training in Chapters 3 and 8.

Definition (Statistical models). A statistical model is a family of probability densities $p(y \mid x, w)$ on $\mathcal{Y}$, one for each input $x \in \mathcal{X}$ and each parameter $w \in W$. If there are no inputs, then we write $p(x \mid w)$ for a family of probability densities on $\mathcal{X}$.

The notion of likelihood was introduced by Fisherⁱⁱ in statistics. Given observations, the likelihood is the probability of the observations, viewed as a function of the parameter. It is not a probability distribution in the parameter; it measures how well each parameter explains what was observed. In training, we can select the optimal parameter by maximizing the likelihood.

Definition (Likelihood). Given a training set $\mathcal{D}_n$, the likelihood of a parameter w is

$$\prod_{j=1}^n p(y_j \mid x_j, w).$$

A maximum likelihood estimator is a parameter $\hat{w}$ which maximizes the likelihood.

Taking $-\frac{1}{n} \log$, maximizing the likelihood is the same as minimizing

$$L_n(w) = -\frac{1}{n} \sum_{j=1}^n \log p(y_j \mid x_j, w).$$

This is the empirical loss for the loss function $\ell = -\log p$, which is called the negative log-likelihood loss, or the cross-entropy loss. Its population version is

$$\bar{L}(w) = \mathbb{E}[-\log p(y \mid x, w)].$$

Remark (The cross-entropy in the scores). For one data point with the scores $f = (f_1, \dots, f_N)$ and the correct output y , the cross-entropy loss is

$$\ell(f, y) = \log \sum_{k=1}^N e^{f_k} - f_y.$$

ⁱJ. Gibbs, *Elementary Principles in Statistical Mechanics*. Charles Scribner’s Sons, 1902.

ⁱⁱR. Fisher, *On the “probable error” of a coefficient of correlation deduced from a small sample*. *Metron* 1 (1921), 3–32; *On the mathematical foundations of theoretical statistics*. *Philos. Trans. Roy. Soc. London Ser. A* 222 (1922), 309–368.

Let $p = (p_1, \dots, p_N)$ be the softmax of f , and let $e_y \in \mathbb{R}^N$ be the coordinate vector of y . Then

$$\nabla_f \ell = p - e_y \quad \text{and} \quad \nabla_f^2 \ell = \text{diag}(p) - pp^\top.$$

The Hessian is the covariance matrix of the coordinate vector e_k when k is drawn from p . So ℓ is convex in the scores. By the chain rule,

$$\nabla L_n(w) = \frac{1}{n} \sum_{j=1}^n Df_w(x_j)^\top (p(\cdot | x_j, w) - e_{y_j}),$$

where $Df_w(x)$ is the Jacobian matrix of $w \mapsto f_w(x)$.

For $N \geq 2$, every softmax probability is less than 1, so the cross-entropy loss L_n is positive. The margin of w on the training set is

$$\gamma(w) = \min_{1 \leq j \leq n} \min_{k \neq y_j} m_{jk}(w) \quad \text{with} \quad m_{jk}(w) = f_w(x_j)_{y_j} - f_w(x_j)_k.$$

It is positive if and only if, at every training point, the correct output has the strictly largest score. Suppose that the model is homogeneous of degree D , i.e.,

$$f_{cw} = c^D f_w \quad \text{for } c > 0.$$

For instance, the quadratic network for modular addition in Chapter 2 is homogeneous of degree 3. Then $L_n(cw)$ is the average over j of

$$\log \left(1 + \sum_{k \neq y_j} e^{-c^D m_{jk}(w)} \right).$$

Each sum in the logarithm is at most $(N-1)e^{-c^D \gamma(w)}$, and for an index j at which $\gamma(w)$ is attained, it is at least $e^{-c^D \gamma(w)}$. If $\gamma(w) \leq 0$, then the term of such an index j is at least $\log 2$. So

$$L_n(cw) \geq \frac{\log 2}{n}.$$

Suppose that $\gamma(w) > 0$. Since

$$\log(1+s) \leq s \quad \text{for } s \geq 0 \quad \text{and} \quad \frac{s}{2} \leq \log(1+s) \quad \text{for } 0 \leq s \leq 1,$$

we obtain

$$\frac{1}{2n} e^{-c^D \gamma(w)} \leq L_n(cw) \leq (N-1) e^{-c^D \gamma(w)}.$$

Hence

$$\left| \frac{1}{c^D} \log \frac{1}{L_n(cw)} - \gamma(w) \right| \leq \frac{\log(nN)}{c^D}.$$

Suppose that γ is positive somewhere. Since the error is uniform on $\{\gamma > 0\}$, minimizing L_n on the sphere $\{|w| = c\}$ for large c amounts to maximizing γ on the unit sphere. This is the low-temperature limit in the remark on Gibbs distributions, with the energies $m_{jk}(w)$ at the temperature c^{-D} . Moreover, if $f_w(x)$ is continuous in w , then L_n has a positive minimum on every closed ball. Hence $|w_t| \rightarrow \infty$ along every trajectory with $L_n(w_t) \rightarrow 0$. This is why the directions $\frac{w_t}{|w_t|}$ and the networks of maximal margin appear in Chapter 12.

1.3. Entropy and the Kullback-Leibler divergence

Why is $-\log p$ the right loss? The answer comes from information theory, founded by Shannonⁱ. In this section, we explain the original meaning of entropy, cross-entropy, and the Kullback-Leibler divergence, and then how they fit into learning.

Definition (Entropy). The entropy of a probability density q on $\mathcal{Y}$ is

$$H(q) = - \int_{\mathcal{Y}} q(y) \log q(y) dy.$$

The quantity $-\log q(y)$ is the *surprise* of the outcome y : a likely outcome is not surprising, and an unlikely one is. The entropy is the average surprise. In Shannon's theory of communication, for a finite set $\mathcal{Y}$, the entropy is the least average number of nats (units of $\log e$) per symbol which is needed to encode long sequences of independent outcomes of q . With $\log_2$ in place of $\log$, it is measured in bits.

Definition (Cross-entropy and Kullback-Leibler divergence). Let q and p be probability densities on $\mathcal{Y}$ with $p > 0$. The cross-entropy of p relative to q is

$$H(q, p) = - \int_{\mathcal{Y}} q(y) \log p(y) dy,$$

and the Kullback-Leibler divergence of p from q is

$$\text{KL}(q||p) = H(q, p) - H(q) = \int_{\mathcal{Y}} q(y) \log \frac{q(y)}{p(y)} dy.$$

The cross-entropy is the average surprise when outcomes are drawn from q but predicted by p , and the Kullback-Leibler divergence is the excess of this surprise over the least possible one. Kullback-Leiblerⁱⁱ introduced it as the mean information for discrimination. For an observation y , the log-likelihood ratio $\log \frac{q(y)}{p(y)}$ is the evidence that y provides for q against p , and $\text{KL}(q||p)$ is the average evidence when q is true. It is not symmetric in q and p , so it is not a metric. The next proposition shows that it is non-negative.

Proposition 1.1 (Gibbs' inequality). *Let q and p be probability densities on $\mathcal{Y}$ with $p > 0$. Then*

$$\text{KL}(q||p) \geq 0,$$

with equality if and only if $p = q$ almost everywhere. Equivalently, $H(q, p) \geq H(q)$, i.e., the average surprise is smallest when the prediction is the truth.

PROOF. Since $\log s \leq s - 1$ for $s > 0$, with equality if and only if $s = 1$,

$$-\text{KL}(q||p) = \int_{\{q>0\}} q \log \frac{p}{q} dy \leq \int_{\{q>0\}} q \left(\frac{p}{q} - 1 \right) dy = \int_{\{q>0\}} p dy - 1 \leq 0,$$

with equality if and only if $p = q$ almost everywhere on $\{q > 0\}$. Since both p and q are probability densities, it follows that $p = q$ almost everywhere. $\square$

We now explain how these notions fit into learning. Suppose that the data are generated by an unknown true density $q_0(y | x)$. Then the population loss of the negative log-likelihood is an average cross-entropy:

$$\bar{L}(w) = \mathbb{E}_x [H(q_0(\cdot | x), p(\cdot | x, w))] = \mathbb{E}_x [H(q_0(\cdot | x))] + \mathbb{E}_x [\text{KL}(q_0(\cdot | x) || p(\cdot | x, w))].$$

ⁱC. Shannon, *A mathematical theory of communication*. Bell System Tech. J. 27 (1948), 379–423, 623–656.

ⁱⁱS. Kullback and R. Leibler, *On information and sufficiency*. Ann. Math. Statistics 22 (1951), 79–86.

The first term does not depend on w ; it is the unavoidable surprise of the data. The second term measures how far the model is from the truth. The model is said to be realizable if

$$q_0 = p(\cdot \mid \cdot, w_0) \quad \text{for some } w_0 \in W.$$

Definition (Normalized loss). We normalize the population loss as

$$L(w) = \bar{L}(w) - \inf_{w' \in W} \bar{L}(w') \geq 0,$$

and we call $W_0 = L^{-1}(0)$ the set of optimal parameters.

Corollary 1.2. *If the model is realizable, then*

$$L(w) = \mathbb{E}_x [\text{KL}(q_0(\cdot \mid x) \| p(\cdot \mid x, w))]$$

and

$$W_0 = \{w \in W : p(\cdot \mid x, w) = q_0(\cdot \mid x) \text{ for almost every } x\}.$$

PROOF. By Proposition 1.1, the second term in the decomposition of $\bar{L}(w)$ is non-negative, and it vanishes at w_0 . So $\inf \bar{L}$ is the first term, and the claims follow from the equality case of Proposition 1.1. $\square$

So minimizing the cross-entropy loss is minimizing the Kullback-Leibler divergence from the truth.

Example (Gaussian regression). Let $p(y \mid x, w)$ be the Gaussian density on $\mathbb{R}^N$ with mean $f_w(x)$ and identity covariance. Then

$$-\log p(y \mid x, w) = \frac{1}{2} |y - f_w(x)|^2 + \frac{N}{2} \log(2\pi).$$

So the negative log-likelihood loss is half the square loss, up to an additive constant. If the truth is $y = f_{w_0}(x) + \xi$ with standard Gaussian noise ξ , then a direct computation with Gaussian densities gives

$$L(w) = \frac{1}{2} \mathbb{E}_x [|f_w(x) - f_{w_0}(x)|^2].$$

So W_0 is the set of parameters which compute the same function as w_0 on the support of the input distribution.

1.4. Fisher information and singular models

The Fisher information measures how much a single observation tells about the parameter. Fisherⁱ introduced it under the name *amount of information*. It is the variance of the score $\nabla_w \log p$, and its inverse bounds how precisely the parameter can be estimated. The next notion separates the classical theory of statistics from the theory needed for neural networks.

Definition (Fisher information). The Fisher information matrix of the model at w is

$$I(w) = \mathbb{E} \left[\nabla_w \log p(y \mid x, w) \nabla_w \log p(y \mid x, w)^\top \right],$$

where x is distributed according to the input distribution and y according to $p(\cdot \mid x, w)$.

ⁱR. Fisher, *Theory of statistical estimation*. Proc. Cambridge Philos. Soc. 22 (1925), 700–725.

Its statistical meaning is the Cramérⁱ-Raoⁱⁱ inequality: in a one-parameter model, every unbiased estimator of w computed from n independent observations has variance at least $\frac{1}{nI(w)}$. So a large Fisher information means that the parameter can be estimated precisely, and a vanishing Fisher information means that a small change of the parameter is invisible in the data.

If differentiation in w commutes with integration in y , then the score has mean zero: for every x ,

$$\int_{\mathcal{Y}} p(y \mid x, w) \nabla_w \log p(y \mid x, w) dy = \nabla_w \int_{\mathcal{Y}} p(y \mid x, w) dy = 0.$$

So $I(w)$ is the covariance matrix of the score.

The Cramér-Rao inequality is proved like the Heisenberg uncertainty principle, with the score in the role of the momentum. Let

$$p_n(w) = \prod_{j=1}^n p(y_j \mid w)$$

be the density of n independent observations. An estimator is a function $T(y_1, \dots, y_n)$, and it is unbiased if

$$\int T p_n(w) dy_1 \cdots dy_n = w \quad \text{for every } w.$$

Since

$$\int \partial_w p_n(w) dy_1 \cdots dy_n = 0,$$

differentiating in w gives

$$\int (T - w) \partial_w \log p_n(w) p_n(w) dy_1 \cdots dy_n = 1.$$

The total score $\partial_w \log p_n$ is the sum of n independent scores with mean zero, so its variance is $nI(w)$. By the Cauchy-Schwarz inequality,

$$1 \leq \text{Var}(T) \cdot nI(w).$$

For $n = 1$ and a location family $p(y \mid w) = g(y - w)$, where g has mean zero, the estimator $T = y_1$ is unbiased, and the inequality is the Heisenberg inequality

$$4 \int_{\mathbb{R}} y^2 \psi^2 dy \int_{\mathbb{R}} \psi'^2 dy \geq 1 \quad \text{for } \psi = \sqrt{g}.$$

Proposition 1.3. *Suppose that the model is realizable, p is C^2 in w , and differentiation in w commutes with integration in y . Then*

$$\nabla^2 L(w_0) = I(w_0).$$

PROOF. By Corollary 1.2,

$$L(w) = \mathbb{E}_x \left[\int_{\mathcal{Y}} q_0 \log q_0 dy - \int_{\mathcal{Y}} q_0 \log p(\cdot \mid x, w) dy \right].$$

So

$$\nabla^2 L(w_0) = -\mathbb{E}_x \left[\int_{\mathcal{Y}} p \nabla^2 \log p dy \right] \quad \text{with } p = p(\cdot \mid x, w_0).$$

ⁱH. Cramér, *Mathematical Methods of Statistics*. Princeton University Press, 1946.

ⁱⁱC. Rao, *Information and the accuracy attainable in the estimation of statistical parameters*. Bull. Calcutta Math. Soc. 37 (1945), 81–91.

Since

$$\nabla^2 \log p = \frac{\nabla^2 p}{p} - \nabla \log p \nabla \log p^t \quad \text{and} \quad \int_{\mathcal{Y}} \nabla^2 p \, dy = \nabla^2 \int_{\mathcal{Y}} p \, dy = 0,$$

we obtain

$$\nabla^2 L(w_0) = \mathbb{E} [\nabla \log p \nabla \log p^t] = I(w_0).$$

□

By Taylor's formula, Proposition 1.3 gives

$$L(w) = \frac{1}{2} (w - w_0)^t I(w_0) (w - w_0) + o(|w - w_0|^2).$$

So the Fisher information is the second-order part of the Kullback-Leibler divergence between nearby distributions.

Definition (Regular and singular models). Suppose that the model is realizable. It is regular if it is identifiable, i.e., $W_0 = \{w_0\}$ consists of one point, and the Fisher information matrix $I(w_0)$ is positive definite. Otherwise, the model is singular.

So the model is regular exactly when L has an isolated non-degenerate minimum, i.e., when L looks like a quadratic bowl near W_0 . In a singular model, different parameters may give the same distribution, or the Fisher information degenerates somewhere on W_0 ; see Watanabe [W].

Example (A coin). Let $w \in (0, 1)$ be the probability that a coin lands heads, so that $p(1 | w) = w$ and $p(0 | w) = 1 - w$, where 1 means heads. The score is

$$\frac{y}{w} - \frac{1 - y}{1 - w},$$

and its variance is

$$I(w) = w \cdot \frac{1}{w^2} + (1 - w) \cdot \frac{1}{(1 - w)^2} = \frac{1}{w(1 - w)} > 0.$$

So the coin is a regular model. If k is the number of heads in n flips, then the frequency $\frac{k}{n}$ is an unbiased estimator of w , and its variance is

$$\frac{w(1 - w)}{n} = \frac{1}{nI(w)}.$$

So equality holds in the Cramér-Rao inequality.

Example (Handwritten digits). The digit network is a singular model: many different parameters compute the same function, as we shall see in Chapter 2. The same holds for every neural network. How singularity changes the theory is the subject of Chapter 6.

1.5. The Bayesian point of view

In the Bayesian point of view, which goes back to Bayesⁱ and Laplaceⁱⁱ, probability measures a degree of belief in an unknown quantity. Beliefs are updated by Bayes' rule when data arrive: for events E and F with $\mathbb{P}(F) > 0$,

$$\mathbb{P}(E | F) = \frac{\mathbb{P}(F | E)\mathbb{P}(E)}{\mathbb{P}(F)}.$$

ⁱT. Bayes, *An essay towards solving a problem in the doctrine of chances*. Philos. Trans. Roy. Soc. London 53 (1763), 370–418.

ⁱⁱP.-S. Laplace, *Mémoire sur la probabilité des causes par les évènements*. Mém. Acad. Roy. Sci. Paris (Savants étrangers) 6 (1774), 621–656.

In learning, the unknown quantity is the parameter w .

Definition (Prior, posterior, and evidence). A prior is a probability density φ on W , which describes our belief before seeing any data. After seeing the training set $\mathcal{D}_n$, Bayes' rule gives the posterior density

$$\pi_n(w) = \frac{\varphi(w) \prod_{j=1}^n p(y_j | x_j, w)}{Z_n}, \quad \text{where } Z_n = \int_W \varphi(w') \prod_{j=1}^n p(y_j | x_j, w') dw'.$$

The number Z_n is the evidence: the probability of the observed outputs, averaged over the prior.

So the posterior is the prior reweighted by the likelihood. Instead of one parameter, the Bayesian point of view keeps all parameters, weighted by how well they explain the data, and predicts by averaging:

$$p(y | x, \mathcal{D}_n) = \int_W p(y | x, w) \pi_n(w) dw.$$

Example (A coin). Take the uniform prior $\varphi \equiv 1$ on $[0, 1]$. If n flips give k heads, then the posterior density is

$$\pi_n(w) = \frac{(n+1)!}{k!(n-k)!} w^k (1-w)^{n-k},$$

by Euler's Beta integral

$$\int_0^1 w^a (1-w)^b dw = \frac{a!b!}{(a+b+1)!}.$$

By the same integral, its mean is $\frac{k+1}{n+2}$ and its variance is

$$\frac{(k+1)(n-k+1)}{(n+2)^2(n+3)} \leq \frac{1}{4(n+3)}.$$

So the posterior concentrates near the observed frequency $\frac{k}{n}$, in a window of width of order $n^{-\frac{1}{2}}$; see Figure 1.2. The posterior mean $\frac{k+1}{n+2}$ is Laplace's rule of succession: after k heads in n flips, the probability of heads at the next flip is $\frac{k+1}{n+2}$.

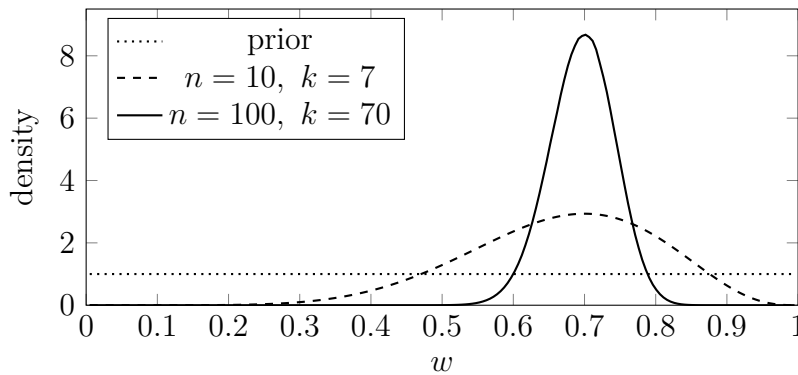


FIGURE 1.2. The uniform prior and the posteriors after 10 and 100 flips of a coin

Remark (The Bernstein-von Mises theorem). In a regular model, the posterior is asymptotically Gaussian: its density is close to the Gaussian density with mean the maximum likelihood

estimator and covariance $(nI(w_0))^{-1}$. This is the Bernstein-von Mises theorem; see Theorem 10.1 of van der Vaartⁱ. It is Laplace's method applied to the integral of

$$\varphi(w) \prod_j p(y_j | x_j, w) = \varphi(w) e^{-nL_n(w)},$$

in which the phase nL_n has a non-degenerate minimum. For the coin, the posterior variance is

$$\frac{w(1-w)}{n} (1 + o(1)) = \frac{1}{nI(w)} (1 + o(1)) \quad \text{with } w = \frac{k}{n},$$

in accordance with the Fisher information computed in Section 1.4.

Example (Handwritten digits). For the digit network, the posterior is a probability density on $\mathbb{R}^{13,002}$, the space of all weights. As n grows, it still concentrates, but it concentrates near the whole set W_0 of optimal parameters, which is not a point. Since the model is singular, the Bernstein-von Mises theorem does not apply, and the posterior is not asymptotically Gaussian. Which parts of W_0 the posterior prefers is the subject of Chapter 6.

Like the softmax in Section 1.2, the posterior is a Gibbs distribution. Indeed, with the negative log-likelihood loss L_n ,

$$\pi_n = \frac{1}{Z_n} \varphi e^{-nL_n}.$$

So the energy is L_n , the reference measure is the prior, and the temperature is $\frac{1}{n}$.

Proposition 1.4 (Gibbs variational principle). *Let $\varphi > 0$ be a probability density on W , let L be a measurable function on W which is bounded below, and let $\varepsilon > 0$. Let*

$$\mu_\varepsilon = \frac{e^{-\frac{L}{\varepsilon}} \varphi}{\int_W e^{-\frac{L}{\varepsilon}} \varphi dw}.$$

Then for every probability density μ on W with $\text{KL}(\mu||\varphi) < \infty$,

$$-\varepsilon \log \int_W e^{-\frac{L}{\varepsilon}} \varphi dw \leq \int_W L \mu dw + \varepsilon \text{KL}(\mu||\varphi),$$

with equality if and only if $\mu = \mu_\varepsilon$ almost everywhere.

PROOF. Let $Z = \int_W e^{-\frac{L}{\varepsilon}} \varphi dw$. Since L is bounded below, $0 < Z < \infty$. If $\int_W L \mu dw = \infty$, then there is nothing to prove. Otherwise, since

$$\log \frac{\mu}{\mu_\varepsilon} = \log \frac{\mu}{\varphi} + \frac{L}{\varepsilon} + \log Z \quad \text{on } \{\mu > 0\},$$

integration against μ gives

$$\varepsilon \text{KL}(\mu||\mu_\varepsilon) = \int_W L \mu dw + \varepsilon \text{KL}(\mu||\varphi) + \varepsilon \log Z.$$

The claim follows from Proposition 1.1. □

If $\varphi > 0$ and L_n is bounded below, for instance if $\mathcal{Y}$ is finite, then Proposition 1.4 applies with $\varepsilon = \frac{1}{n}$ and $L = L_n$. The minimizer is the posterior π_n , and the minimum is $-\frac{1}{n} \log Z_n$. So the posterior balances the fit to the data, $\int_W L_n \mu dw$, against the cost $\frac{1}{n} \text{KL}(\mu||\varphi)$ of moving away from the prior. If φ is uniform on W , then

$$\text{KL}(\mu||\varphi) = \log \text{Vol}(W) - H(\mu).$$

ⁱA. van der Vaart, *Asymptotic Statistics*. Cambridge University Press, 1998.

So, up to the constant $\frac{1}{n} \log \text{Vol}(W)$, the posterior minimizes the mean energy minus the temperature times the entropy, and the minimum is the free energy of statistical mechanics. Divided by the temperature $\frac{1}{n}$, the minimum $-\frac{1}{n} \log Z_n$ becomes

$$F_n = -\log Z_n.$$

This is why F_n is called the free energy in Chapter 6.

The volume enters when μ is the normalized restriction of φ to a sublevel set. Suppose that $L \geq 0$, and let

$$V(t) = \int_{\{L < t\}} \varphi dw \quad \text{for } t > 0.$$

If $V(t) > 0$, then the density $\mu = \frac{1}{V(t)} \varphi \mathbb{1}_{\{L < t\}}$ satisfies

$$\int_W L \mu dw \leq t \quad \text{and} \quad \text{KL}(\mu \| \varphi) = \log \frac{1}{V(t)}.$$

Hence

$$-\varepsilon \log \int_W e^{-\frac{L}{\varepsilon}} \varphi dw \leq \inf_{t>0} \left(t + \varepsilon \log \frac{1}{V(t)} \right).$$

The first term is the depth of the loss, and the second one is the price of a small volume. If

$$V(t) = ct^\lambda (1 + o(1)) \quad \text{as } t \rightarrow 0^+$$

for some $c, \lambda > 0$, then the choice $t = \lambda \varepsilon$ gives

$$-\varepsilon \log \int_W e^{-\frac{L}{\varepsilon}} \varphi dw \leq \lambda \varepsilon \log \frac{1}{\varepsilon} + O(\varepsilon) \quad \text{as } \varepsilon \rightarrow 0^+.$$

By Proposition 6.2 with $n = \frac{1}{\varepsilon}$ and $m = 1$, this bound is sharp up to $O(\varepsilon)$. This trade-off between depth and volume is behind the learning coefficient in Chapter 6, the effective losses in Chapter 7, and the entropic selection in Chapter 10.

1.6. Generalization

A model generalizes if it predicts well on new data from the same source, i.e., if its population loss is small and not only its empirical loss. Recall that the empirical loss is computed on the training data, while the population loss is the average over all possible data. In practice, one estimates the population loss on a held-out test set, a second set of samples that is not used in training. Since the test data are independent of the trained parameter, the test loss is an unbiased estimate of its population loss. By the law of large numbers, it converges to the population loss as the size of the test set grows.

Classically, a model with as many parameters as data points is expected to overfit, i.e., to fit the noise in the training data, so that its empirical loss is much smaller than its population loss. Modern neural networks do not fit this picture. They have far more parameters than data points, they reach zero empirical loss, and they still generalize. Moreover, large networks can reach zero empirical loss even when the labels are replaced by pure noise, as observed by Zhang et al.ⁱ

Example (Handwritten digits with random labels). Replace the labels of the 60,000 training images by independent uniformly random digits. Suppose that a large network still reaches zero empirical loss on them. Since the labels of new images are then independent of the images, every model predicts a new label correctly with probability $\frac{1}{10}$, so the network has learned nothing that

ⁱC. Zhang, S. Bengio, M. Hardt, et al., *Understanding deep learning requires rethinking generalization*. International Conference on Learning Representations, 2017.

generalizes. The same architecture, trained on the true labels, generalizes well. So the number of parameters alone cannot explain generalization: what matters is which of the many parameters with zero empirical loss training finds. This question is a central theme of Part III.

There are two regimes of generalization, and the distinction is crucial for safety. Generalization *in distribution* means that the new inputs are drawn from the same distribution as the training data; this is what a held-out test set measures. Generalization *out of distribution* means that the new inputs come from a different distribution, for example because the world has changed after the system is deployed; no test set drawn from the training distribution can measure this. We return to this distinction in Chapter 4.

CHAPTER 2

Neural networks

A *neural network* is a model whose functions f_w are compositions of simple maps, called layers. The parameters, called weights, are the entries of the matrices and vectors in the layers. In this chapter, we define neural networks, describe the digit network and the GPT models, give the examples which appear in Part III, and explain why neural networks are singular models.

2.1. Layers and neurons

Definition (Layers). Let $\sigma : \mathbb{R} \rightarrow \mathbb{R}$ be a fixed function, called the activation function. A layer is a map

$$\mathbb{R}^M \rightarrow \mathbb{R}^N, \quad x \mapsto \sigma(Ax + b),$$

where $A \in M_{N,M}(\mathbb{R})$, $b \in \mathbb{R}^N$, and σ is applied to each coordinate. Let a_j be the j -th row of A . The j -th coordinate $x \mapsto \sigma(\langle a_j, x \rangle + b_j)$ is called the j -th neuron of the layer. The matrix A and the vector b are called the weights and the biases of the layer.

A neuron adds up its inputs with the weights a_j , shifts the sum by the bias b_j , and then “fires” according to σ .

Example (Activation functions).

- (1) [Sigmoid] The sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

increases from 0 to 1; see Figure 2.1. It is a smooth version of the step function $\mathbf{1}_{\{z>0\}}$: a neuron with this activation is nearly off for $z \ll 0$ and nearly fully on for $z \gg 0$. It is the activation function of the digit network below.

- (2) [ReLU] The rectified linear unit

$$\text{ReLU}(z) = \max\{0, z\}$$

is the common choice in modern networks. It is piecewise linear.

- (3) [Quadratic] The quadratic activation $\sigma(z) = z^2$ makes f_w a polynomial in w and x , which is convenient for algebraic geometry.
- (4) [Identity] The identity $\sigma(z) = z$ gives a linear layer.

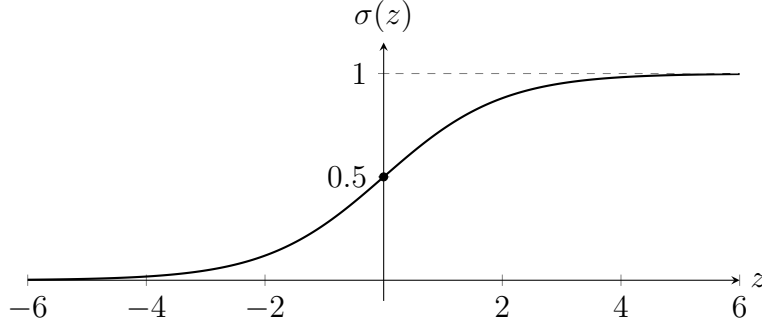


FIGURE 2.1. The sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$, which increases from 0 to 1 and takes the value $\frac{1}{2}$ at $z = 0$

Definition (Neural networks). A neural network of depth D is a composition

$$f_w = \ell_D \circ \ell_{D-1} \circ \cdots \circ \ell_1$$

of layers $\ell_k(x) = \sigma(A_k x + b_k)$, except that the last layer $\ell_D(x) = A_D x + b_D$ is often affine. The parameter w consists of all entries of $A_1, b_1, \dots, A_D, b_D$, which are also called the weights. The output dimension of an intermediate layer is called its width.

The values that flow through the network on a given input, i.e., the outputs of the intermediate layers, are called activations. They are the internal “thoughts” of the network, and they are the main object of study in interpretability (Chapter 4).

Example (Handwritten digits). The digit network has depth 3:

$$f_w = \ell_3 \circ \ell_2 \circ \ell_1 : \mathbb{R}^{784} \xrightarrow{\ell_1} \mathbb{R}^{16} \xrightarrow{\ell_2} \mathbb{R}^{16} \xrightarrow{\ell_3} \mathbb{R}^{10},$$

with the sigmoid activation in every layer. The 784 input coordinates are the pixels, the two hidden layers have 16 neurons each, and the 10 outputs are the scores of the digits; see Figure 2.2. The number of parameters is

$$d = \underbrace{784 \cdot 16 + 16}_{\ell_1} + \underbrace{16 \cdot 16 + 16}_{\ell_2} + \underbrace{16 \cdot 10 + 10}_{\ell_3} = 13,002.$$

One may hope that the first hidden layer detects small pieces of strokes, the second one combines them into strokes and loops, and the output layer combines those into digits. Whether a trained network actually computes in this way is a question of interpretability (Chapter 4).

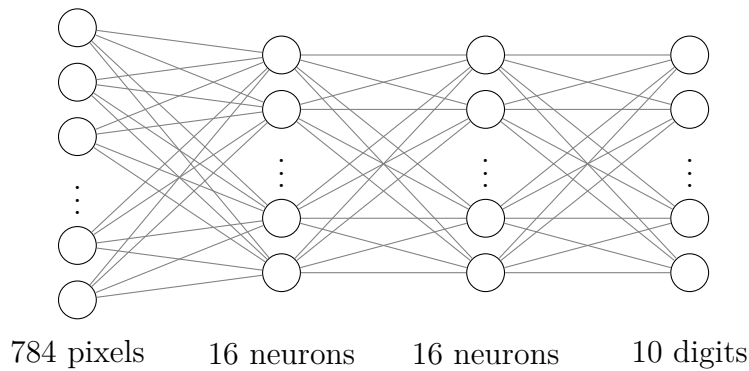


FIGURE 2.2. The digit network, with the input layer, two hidden layers, and the output layer

Example.

- (1) [Two-layer networks] A two-layer network with width H and scalar output is

$$f_w(x) = \sum_{j=1}^H a_j \sigma(\langle u_j, x \rangle + b_j) \quad \text{with } w = (a_j, u_j, b_j)_{j=1}^H \in \mathbb{R}^{H(M+2)}.$$

Each summand is one hidden neuron.

- (2) [Two-layer linear networks] With $\sigma(z) = z$ and no biases, a two-layer network is the map $x \mapsto BAx$ with $A \in M_{H,M}(\mathbb{R})$ and $B \in M_{N,H}(\mathbb{R})$. The function it computes is linear, but the parametrization by the pair (A, B) is not. This model is the subject of Section 3.5.
- (3) [A quadratic network for modular addition] Let p be an odd integer. On inputs $(a, b) \in (\mathbb{Z}/p\mathbb{Z})^2$, consider

$$f_w(a, b) = \sum_{j=1}^H V_j \left(u'_j(a) + u''_j(b) \right)^2 \in \mathbb{R}^p,$$

where $u'_j, u''_j : \mathbb{Z}/p\mathbb{Z} \rightarrow \mathbb{R}$ are lookup tables and $V_j \in \mathbb{R}^p$. It is trained to output the coordinate vector indexed by $a + b$. Gromovⁱ gave explicit weights with which such a network solves modular addition. He also observed grokking (Section 3.4) in its training. We return to this network in Chapter 12.

2.2. Symmetries and singularities

Different weights can compute the same function. This simple fact makes neural networks singular models.

Example (Handwritten digits). In the digit network, permuting the 16 neurons of a hidden layer, together with their incoming and outgoing weights, does not change f_w . So every function computed by the digit network is computed by at least $(16!)^2 \approx 4.4 \times 10^{26}$ parameters, if the neurons in each hidden layer have distinct weights. These parameters form a finite set. Continuous symmetries give positive-dimensional sets, as follows.

Proposition 2.1. *Let $A \in M_{H,M}(\mathbb{R})$ and $B \in M_{N,H}(\mathbb{R})$. For every $G \in GL(H, \mathbb{R})$,*

$$(BG^{-1})(GA) = BA.$$

If, moreover, $\text{rank } A = H$, then $GA = A$ implies $G = I$. So the orbit

$$\{(GA, BG^{-1}) : G \in GL(H, \mathbb{R})\}$$

is an immersed submanifold of dimension H^2 on which the function $x \mapsto BAx$ is constant.

PROOF. The identity is clear. If $\text{rank } A = H$, then A has a right inverse A^+ with $AA^+ = I$. So $GA = A$ gives

$$G = GAA^+ = AA^+ = I.$$

Hence the map $G \mapsto (GA, BG^{-1})$ is injective. Its differential at G maps X to $(XA, -BG^{-1}XG^{-1})$, which vanishes only if $XA = 0$, i.e., $X = 0$. So it is an injective immersion of the H^2 -dimensional group $GL(H, \mathbb{R})$. $\square$

Remark (Further symmetries). If $a_j = 0$, then the incoming weights (u_j, b_j) of the j -th neuron can be changed freely. More generally, if a column of one weight matrix is zero, then the corresponding row of the matrix before it is free.

ⁱA. Gromov, *Grokking modular arithmetic*. arXiv:2301.02679.

Consequently, the set W_0 of optimal parameters of a neural network is typically not a point but a positive-dimensional set, and the Hessian of L is degenerate along it. Moreover, W_0 is usually not a manifold: several families of optimal parameters meet along singularities.

Example (The smallest singular network). Let $M = N = H = 1$ in the two-layer linear network. Suppose that the target is the zero function. With whitened inputs (see Section 3.5), the loss is

$$L(a, b) = \frac{1}{2}(ba)^2 \quad \text{for } (a, b) \in \mathbb{R}^2.$$

Its zero set $W_0 = \{ab = 0\}$ is the union of the two coordinate axes, which cross at a singular point, the origin; see Figure 2.3. The Hessian of L vanishes at the origin, and it has rank one at the other points of W_0 .

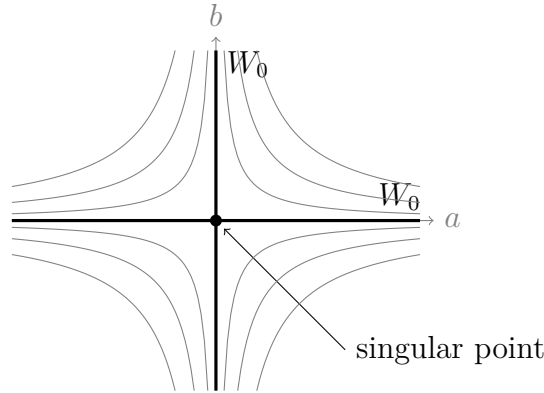


FIGURE 2.3. The zero set of $L(a, b) = \frac{1}{2}(ba)^2$ and some level curves of L

2.3. Language models and agents

The AI systems that motivate AI safety are mainly of two kinds. We describe them only to the extent needed in these notes; the terms are collected in the glossary, Appendix A.

A *language model* is a neural network trained to predict the next *token* (a word or a piece of a word) of a text, with the cross-entropy loss of Section 1.2. The *transformer*, introduced by Vaswani et al.ⁱ, is the dominant architecture of language models. It alternates layers as above with attention layers, which mix information across the positions of the text.

An *agent* acts in an environment. Its decision rule, a possibly randomized map from observations to actions, is called a *policy*. In *reinforcement learning*, the actions are scored by a numerical reward, and the policy, given by a neural network, is trained to earn more reward.

Example (Language models). The GPT models are language models: GPT stands for “generative pre-trained transformer”. Chat systems such as ChatGPT are built on GPT models which are further trained to follow instructions; see Ouyang et al.ⁱⁱ. They increasingly act as agents, for example by writing and running computer programs.

ⁱA. Vaswani, N. Shazeer, N. Parmar, et al., *Attention is all you need*. Advances in Neural Information Processing Systems 30 (2017), 5998–6008.

ⁱⁱL. Ouyang, J. Wu, X. Jiang, et al., *Training language models to follow instructions with human feedback*. Advances in Neural Information Processing Systems 35 (2022), 27730–27744.

CHAPTER 3

Training

Training is the process that selects the parameters of a model from data. It is an iterative algorithm, which moves the parameter downhill on the empirical loss. In this chapter, we introduce gradient descent, stochastic gradient descent, and the Langevin dynamics, which is the continuous-time model of noisy training used in Part III. We end with the deep linear network, in which training can be solved explicitly.

3.1. Gradient descent and gradient flow

Let $L : \mathbb{R}^d \rightarrow [0, \infty)$ be a C^1 loss (empirical or population).

Definition (Gradient descent and gradient flow). Let $\eta > 0$. Gradient descent with learning rate η is the iteration

$$w_{k+1} = w_k - \eta \nabla L(w_k) \quad \text{for } k = 0, 1, 2, \dots$$

Gradient flow is its continuous-time limit $\eta \rightarrow 0$, i.e., the ordinary differential equation

$$\dot{w}(t) = -\nabla L(w(t)).$$

Since $-\nabla L(w)$ is the direction in which L decreases fastest, gradient descent walks downhill on the graph of L .

By the chain rule, along gradient flow,

$$\frac{d}{dt}L(w(t)) = \langle \nabla L(w(t)), \dot{w}(t) \rangle = -|\nabla L(w(t))|^2 \leq 0.$$

So gradient flow decreases the loss until it reaches a critical point, $\nabla L(w) = 0$. For real-analytic L , every bounded trajectory of gradient flow converges to a single critical point, by a theorem of Łojasiewicz¹. The critical point may be a minimum, or a saddle point, at which L decreases in some directions. Gradient flow from a generic initial point does not end at a saddle point, but it may spend a long time near one. This is the origin of the plateaus in Section 3.4.

Remark (The Łojasiewicz inequality). The theorem of Łojasiewicz rests on his gradient inequality. If L is real-analytic near a point w_* , then there are $\theta \in [\frac{1}{2}, 1)$, $c > 0$, and a ball U around w_* such that

$$|\nabla L(w)| \geq c |L(w) - L(w_*)|^\theta \quad \text{for } w \in U.$$

This is clear if $\nabla L(w_*) \neq 0$, and $\theta = \frac{1}{2}$ works at a non-degenerate critical point. For $L(w) = w^{2k}$ on $\mathbb{R}$,

$$|L'(w)| = 2kL(w)^{1-\frac{1}{2k}},$$

so the best exponent at 0 is $\theta = 1 - \frac{1}{2k}$. For $L(a, b) = a^2b^2$ on $\mathbb{R}^2$,

$$|\nabla L(a, b)|^2 = 4a^2b^2(a^2 + b^2) \geq 8|ab|^3,$$

¹S. Łojasiewicz, *Sur les trajectoires du gradient d'une fonction analytique*. Seminari di Geometria 1982–1983, Università di Bologna, 1984, 115–117.

with equality on the diagonals $a = \pm b$, so the best exponent at 0 is $\theta = \frac{3}{4}$. The learning coefficients of these two examples are $\lambda = \frac{1}{2k}$ and $\lambda = \frac{1}{2}$ (Chapter 6). So $\theta = 1 - \lambda$ in the first example. For $k = 1$, however, the two examples have the same $\lambda = \frac{1}{2}$ but the exponents $\theta = \frac{1}{2}$ and $\theta = \frac{3}{4}$. Like λ , the exponent θ measures the flatness of L at w_* , but it is a different invariant.

For the convergence, let $w(t)$ be a trajectory of gradient flow with an accumulation point w_* . Then $L(w(t))$ decreases to $L(w_*)$. Replacing L by $L - L(w_*)$, we may assume that $L(w_*) = 0 < L(w(t))$ for all t ; otherwise, the trajectory is constant after some time. While $w(t) \in U$,

$$\frac{d}{dt} L(w(t))^{1-\theta} = -(1-\theta) L^{-\theta} |\nabla L|^2 \leq -c(1-\theta) |\nabla L| = -c(1-\theta) |\dot{w}(t)|.$$

So, if $w(t) \in U$ for $t_0 \leq t \leq t_1$, then the length of the trajectory between the times t_0 and t_1 satisfies

$$\int_{t_0}^{t_1} |\dot{w}(t)| dt \leq \frac{L(w(t_0))^{1-\theta}}{c(1-\theta)}.$$

Choose t_0 such that $|w(t_0) - w_*|$ and the right-hand side are both smaller than half the radius of U . Then the trajectory never leaves U after the time t_0 . Hence it has finite length, and it converges to w_* . Problem 10.1(c) asks for a stochastic version of this convergence, and Problem 12.3(2) asks for the convergence of the directions $\frac{w(t)}{|w(t)|}$ when $|w(t)| \rightarrow \infty$.

3.2. Stochastic gradient descent

For a large training set, computing ∇L_n is expensive, since it is a sum of n terms. Stochastic gradient descent replaces it by a cheap random estimate. It goes back to the stochastic approximation method of Robbins-Monroⁱ.

Definition (Stochastic gradient descent). Write

$$L_n = \frac{1}{n} \sum_{j=1}^n \ell_j \quad \text{with } \ell_j(w) = \ell(f_w(x_j), y_j).$$

Let $b \leq n$. At each step k , choose a subset $S_k \subset \{1, \dots, n\}$ of size b , called a minibatch, uniformly at random and independently of the past, and set

$$w_{k+1} = w_k - \eta g_k, \quad \text{where } g_k = \frac{1}{b} \sum_{j \in S_k} \nabla \ell_j(w_k).$$

Proposition 3.1. *The estimate g_k is unbiased, that is,*

$$\mathbb{E}[g_k \mid w_k] = \nabla L_n(w_k).$$

PROOF. For a uniformly random subset S_k of size b , each index j belongs to S_k with probability $\frac{b}{n}$. Hence

$$\mathbb{E}[g_k \mid w_k] = \frac{1}{b} \sum_{j=1}^n \mathbb{P}(j \in S_k) \nabla \ell_j(w_k) = \frac{1}{b} \cdot \frac{b}{n} \sum_{j=1}^n \nabla \ell_j(w_k) = \nabla L_n(w_k).$$

□

So stochastic gradient descent (SGD) is gradient descent plus a noise of mean zero; see Figure 3.1. That is,

$$w_{k+1} = w_k - \eta g_k = w_k - \eta \nabla L_n(w_k) - \eta \xi_k \quad \text{with } \xi_k = g_k - \nabla L_n(w_k).$$

ⁱH. Robbins and S. Monro, *A stochastic approximation method*. Ann. Math. Statistics 22 (1951), 400–407.

The covariance of the noise depends on w_k . Suppose that the model fits every data point exactly at w , i.e., $\ell_j(w) = 0$ for all j . Since each $\ell_j \geq 0$ attains its minimum at w , we have $\nabla \ell_j(w) = 0$ for all j . So the noise vanishes at w . In this sense, the noise of SGD typically dies on the set of optimal parameters.

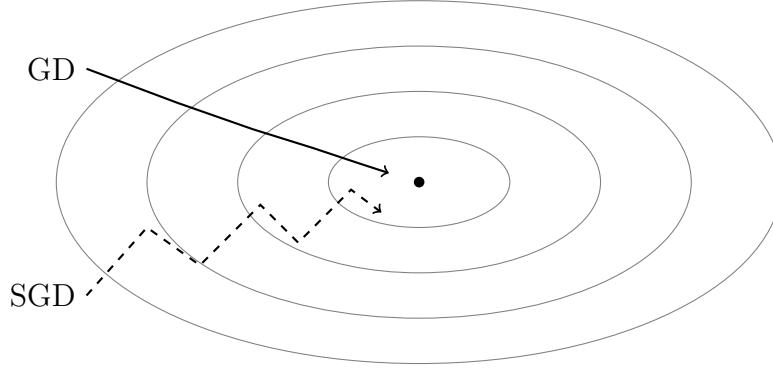


FIGURE 3.1. Gradient descent (solid) and stochastic gradient descent (dashed) on the level sets of a loss

Proposition 3.2. *Let $n \geq 2$. The covariance matrix of the noise is*

$$\Sigma(w_k) = \mathbb{E} [\xi_k \xi_k^t \mid w_k] = \frac{n-b}{b(n-1)} \Sigma_1(w_k),$$

where

$$\Sigma_1(w) = \frac{1}{n} \sum_{j=1}^n (\nabla \ell_j(w) - \nabla L_n(w)) (\nabla \ell_j(w) - \nabla L_n(w))^t$$

is the covariance matrix of the gradients of the individual losses.

PROOF. Let $v_j = \nabla \ell_j(w_k) - \nabla L_n(w_k)$. Then

$$\sum_{j=1}^n v_j = 0 \quad \text{and} \quad \xi_k = \frac{1}{b} \sum_{j \in S_k} v_j.$$

Two distinct indices $i \neq j$ both belong to S_k with probability $\frac{b(b-1)}{n(n-1)}$. Hence

$$\mathbb{E} [\xi_k \xi_k^t \mid w_k] = \frac{1}{b^2} \left(\frac{b}{n} \sum_{j=1}^n v_j v_j^t + \frac{b(b-1)}{n(n-1)} \sum_{i \neq j} v_i v_j^t \right).$$

Since

$$\sum_{j=1}^n v_j v_j^t = n \Sigma_1(w_k) \quad \text{and} \quad \sum_{i \neq j} v_i v_j^t = \left(\sum_{i=1}^n v_i \right) \left(\sum_{j=1}^n v_j \right)^t - \sum_{j=1}^n v_j v_j^t = -n \Sigma_1(w_k),$$

we obtain

$$\mathbb{E} [\xi_k \xi_k^t \mid w_k] = \frac{1}{b} \left(1 - \frac{b-1}{n-1} \right) \Sigma_1(w_k) = \frac{n-b}{b(n-1)} \Sigma_1(w_k).$$

□

So the noise is at most of order $b^{-\frac{1}{2}}$, and of this order if $b \leq \frac{n}{2}$. Its shape is the covariance of the individual gradients. Two cases are typical.

- (1) *Exact fits*: Consider the square loss $\ell_j(w) = \frac{1}{2}|f_w(x_j) - y_j|^2$. Let $J_j(w)$ be the Jacobian matrix of $w \mapsto f_w(x_j)$. Then

$$\nabla \ell_j(w) = J_j(w)^t (f_w(x_j) - y_j).$$

Since $\frac{n-b}{n-1} \leq 1$ and a covariance matrix is at most the matrix of second moments,

$$\text{Tr } \Sigma(w) \leq \frac{1}{bn} \sum_{j=1}^n |\nabla \ell_j(w)|^2 \leq \frac{2}{b} \max_{1 \leq j \leq n} \|J_j(w)\|^2 L_n(w).$$

So, where the Jacobians are bounded, the size of the noise is at most a multiple of $L_n^{\frac{1}{2}}$. This motivates the degenerate noise $\sqrt{2\varepsilon L} dB_t$ in Problem 10.1(c).

- (2) *Noisy labels*: Consider the negative log-likelihood loss $\ell_j(w) = -\log p(y_j | x_j, w)$ of a realizable model, and let $w_0 \in W_0$. Since the vectors $-\nabla \ell_j(w_0)$ are independent samples of the score of Section 1.4, whose covariance matrix is $I(w_0)$, the law of large numbers and Proposition 1.3 give

$$\Sigma(w_0) \approx \frac{1}{b} I(w_0) = \frac{1}{b} \nabla^2 L(w_0) \quad \text{for large } n \text{ and } b \ll n.$$

Suppose that W_0 is smooth near w_0 . Since $L \geq 0$ vanishes on W_0 ,

$$v^t I(w_0) v = v^t \nabla^2 L(w_0) v = 0$$

for every tangent vector v of W_0 at w_0 . So the score is orthogonal to v almost surely, and the noise has no component tangent to W_0 .

3.3. The Langevin dynamics as a model of noisy training

Write SGD as

$$w_{k+1} - w_k = -\eta \nabla L_n(w_k) - \eta \xi_k.$$

Over K steps, the drift contributes about $-K\eta \nabla L_n$, i.e., it moves the parameter along gradient flow for the time $t = K\eta$. The noise contributes a sum of K terms of mean zero and size η , whose variance is of order $K\eta^2 = t\eta$. So on the time scale of gradient flow, the noise looks like a Brownian motion (see Chapter 8) whose variance is proportional to η . This heuristic motivates the following model, named after Langevinⁱ.

Definition (Langevin dynamics). Let $\varepsilon > 0$. The overdamped Langevin dynamics at temperature ε is the stochastic differential equation

$$dw_t = -\nabla L(w_t) dt + \sqrt{2\varepsilon} dB_t,$$

where B_t is a standard Brownian motion in $\mathbb{R}^d$.

We make this definition precise in Chapter 8. Its stationary distribution is the Gibbs measure with density proportional to $e^{-\frac{L(w)}{\varepsilon}}$, which concentrates on the low values of L as $\varepsilon \rightarrow 0$. The Langevin dynamics is the mathematician's proxy for SGD: it is isotropic and reversible, and it has an explicit stationary law. Genuine SGD noise is none of these things. Chaudhari-Soattoⁱⁱ observed that its covariance is highly anisotropic and of low rank, and that the most likely trajectories of SGD resemble closed loops, instead of converging to critical points. In Part III, we study the Langevin dynamics, and we indicate where the difference matters (Chapter 10).

ⁱP. Langevin, *Sur la théorie du mouvement brownien*. C. R. Acad. Sci. Paris 146 (1908), 530–533.

ⁱⁱP. Chaudhari and S. Soatto, *Stochastic gradient descent performs variational inference, converges to limit cycles for deep networks*. International Conference on Learning Representations, 2018.

Remark (The diffusion approximation). The heuristic at the beginning of this section can be made precise in the language of semigroups. Consider the iteration

$$w_{k+1} = w_k - \eta \nabla L(w_k) - \eta \xi_k,$$

where, given $w_k = w$, the noise ξ_k has mean zero, covariance matrix $\Sigma(w)$, and bounded third moments. Let f be a smooth test function. Suppose also that f , ∇L , and Σ are bounded with all their derivatives. By Taylor's formula,

$$\mathbb{E}[f(w_{k+1}) \mid w_k = w] = f - \eta \nabla L \cdot \nabla f + \frac{\eta^2}{2} \left(\text{Tr}(\Sigma \nabla^2 f) + \nabla L^\top \nabla^2 f \nabla L \right) + O(\eta^3),$$

where the functions on the right-hand side are evaluated at w . Let

$$\mathcal{L}^\eta = -\nabla \tilde{L} \cdot \nabla + \frac{\eta}{2} \sum_{i,j=1}^d \Sigma_{ij} \partial_i \partial_j \quad \text{with } \tilde{L} = L + \frac{\eta}{4} |\nabla L|^2.$$

Since $\nabla |\nabla L|^2 = 2 \nabla^2 L \nabla L$, a direct computation gives

$$\mathbb{E}[f(w_{k+1}) \mid w_k = w] = f + \eta \mathcal{L}^\eta f + \frac{\eta^2}{2} (\mathcal{L}^\eta)^2 f + O(\eta^3).$$

So one step of SGD acts on test functions as the semigroup $e^{t\mathcal{L}^\eta}$ at the time $t = \eta$, up to $O(\eta^3)$. With bounds on the derivatives of $e^{t\mathcal{L}^\eta} f$, these errors add up to $O(\eta^2)$ over $\frac{1}{\eta}$ steps. By Itô's formula (Chapter 8), with the diffusion matrix $\eta \Sigma$ in place of $2\varepsilon \text{I}$, the operator $\mathcal{L}^\eta$ is the generator of the diffusion

$$dw_t = -\nabla \tilde{L}(w_t) dt + (\eta \Sigma(w_t))^{\frac{1}{2}} dB_t.$$

The noise enters the generator at the relative order η , the same as the correction $\frac{\eta}{4} |\nabla L|^2$, which comes from the discreteness of the steps. The noise dominates near the critical points of L , where ∇L is small, and it decides the behavior over long times, such as the escape from wells and saddles. If $\Sigma = \sigma^2 \text{I}$, then the diffusion is the Langevin dynamics, with $\tilde{L}$ in place of L , at the temperature

$$\varepsilon = \frac{\eta \sigma^2}{2}.$$

By Proposition 3.2, the temperature of SGD is of order $\frac{\eta}{b}$. In general, it is the matrix $\frac{\eta}{2} \Sigma(w)$.

Remark (Langevin dynamics as a sampler). The Langevin dynamics is also used in practice for a different purpose. Let φ be a prior density. With L replaced by $nL_n - \log \varphi$ and with $\varepsilon = 1$, the stationary density of the Langevin dynamics is proportional to φe^{-nL_n} , which is the Bayesian posterior of Section 1.5. Run in discrete time with minibatch gradients, this dynamics is SGD with Gaussian noise added at each step. It is used to sample from the posterior, as proposed by Welling-Tehⁱ.

3.4. Plateaus and stages

A plot of the loss against training time is called a training curve. Training curves often do not decrease smoothly. Long flat stretches, called plateaus, give way to sudden drops. Each drop often coincides with the formation of a new piece of internal structure in the network, such as the induction heads of Olsson et al.ⁱⁱ; see also [HWFR]. A related phenomenon is *grokking*,

ⁱM. Welling and Y. Teh, *Bayesian learning via stochastic gradient Langevin dynamics*. Proceedings of the 28th International Conference on Machine Learning (2011), 681–688.

ⁱⁱC. Olsson, N. Elhage, N. Nanda, et al., *In-context learning and induction heads*. arXiv:2209.11895.

observed by Power et al.ⁱ: a network fits its training data early, but generalizes only much later, and abruptly. The next section gives the one model in which plateaus can be computed exactly.

3.5. The deep linear network

Fix integers $M, N, H \geq 1$ and a target matrix $\Phi \in M_{N,M}(\mathbb{R})$ of rank $r \leq H$, with the singular value decomposition

$$\Phi = \sum_{\alpha=1}^r s_{\alpha} u_{\alpha} v_{\alpha}^{\text{t}} \quad \text{with } s_1 > s_2 > \cdots > s_r > 0,$$

where $\{u_{\alpha}\}$ and $\{v_{\alpha}\}$ are orthonormal in $\mathbb{R}^N$ and $\mathbb{R}^M$, respectively. Suppose that the data satisfy $y = \Phi x + \xi$, where the noise ξ is independent of x and has mean zero. Suppose also that the inputs are whitened, i.e., $\mathbb{E}[xx^{\text{t}}] = I$. For the two-layer linear network $x \mapsto BAx$,

$$\begin{aligned} \mathbb{E} \left[\frac{1}{2} |y - BAx|^2 \right] &= \frac{1}{2} \mathbb{E} \left[\text{Tr} \left((BA - \Phi) x x^{\text{t}} (BA - \Phi)^{\text{t}} \right) \right] + \frac{1}{2} \mathbb{E} |\xi|^2 \\ &= \frac{1}{2} \|BA - \Phi\|_F^2 + \frac{1}{2} \mathbb{E} |\xi|^2, \end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm. So, up to a constant, the population loss is

$$L(A, B) = \frac{1}{2} \|BA - \Phi\|_F^2 \quad \text{for } A \in M_{H,M}(\mathbb{R}) \text{ and } B \in M_{N,H}(\mathbb{R}). \quad (3.1)$$

It is a polynomial of degree four in the weights, and it is not convex, because the parametrization is a product. Its set of optimal parameters is the fiber

$$W_0 = F_{\Phi} = \{(A, B) : BA = \Phi\}.$$

The critical points of (3.1) are classified by Baldi-Hornikⁱⁱ and by Zhou-Liangⁱⁱⁱ. They occur where, for a subset $S \subset \{1, \dots, r\}$ of the modes,

$$BA = \sum_{\alpha \in S} s_{\alpha} u_{\alpha} v_{\alpha}^{\text{t}} \quad \text{and} \quad Av_{\alpha} = 0, \quad B^{\text{t}}u_{\alpha} = 0 \quad \text{for } \alpha \notin S.$$

Every critical point which is not a minimum is a saddle point. We write C_k for the critical set with $S = \{1, \dots, k\}$, at which only the k largest modes are learned, and

$$L_k = \frac{1}{2} \sum_{\alpha > k} s_{\alpha}^2 \quad (3.2)$$

for its loss. Then $L_0 > L_1 > \cdots > L_r = 0$, and $C_r = W_0$.

The classification follows from the formulas

$$\partial_A L = B^{\text{t}}(BA - \Phi) \quad \text{and} \quad \partial_B L = (BA - \Phi)A^{\text{t}}.$$

Fix a critical point (A, B) , and write $P = BA$ and $E = \Phi - P$. Then $B^{\text{t}}E = 0$ and $EA^{\text{t}} = 0$. Hence

$$P^{\text{t}}E = A^{\text{t}}B^{\text{t}}E = 0 \quad \text{and} \quad PE^{\text{t}} = B(EA^{\text{t}})^{\text{t}} = 0.$$

So P and E have orthogonal column spaces and orthogonal row spaces, and

$$\Phi\Phi^{\text{t}} = PP^{\text{t}} + EE^{\text{t}}.$$

ⁱA. Power, Y. Burda, H. Edwards, et al., *Grokking: generalization beyond overfitting on small algorithmic datasets*. arXiv:2201.02177.

ⁱⁱP. Baldi and K. Hornik, *Neural networks and principal component analysis: learning from examples without local minima*. Neural Networks 2 (1989), 53–58.

ⁱⁱⁱY. Zhou and Y. Liang, *Critical points of linear neural networks: analytical forms and landscape properties*. International Conference on Learning Representations, 2018.

This matrix preserves the column space of P and its orthogonal complement. Since its non-zero eigenvalues s_α^2 are simple, each u_α lies in one of these two spaces. Let S be the set of α with u_α in the column space of P . Then $P^t u_\alpha = \Phi^t u_\alpha = s_\alpha v_\alpha$ for $\alpha \in S$, and $P^t u_\alpha = 0$ for $\alpha \notin S$. Since $\Phi\Phi^t \geq PP^t$, the column space of P lies in that of Φ , which is spanned by the u_α . So

$$P = \sum_{\alpha \in S} s_\alpha u_\alpha v_\alpha^t \quad \text{and} \quad E = \sum_{\alpha \notin S} s_\alpha u_\alpha v_\alpha^t.$$

Finally, $B^t E = 0$ and $EA^t = 0$ give $B^t u_\alpha = 0$ and $Av_\alpha = 0$ for $\alpha \notin S$. Conversely, these conditions give $B^t E = 0$ and $EA^t = 0$, so (A, B) is a critical point.

The saddle points are strict, i.e., at every critical point with $S \neq \{1, \dots, r\}$, the Hessian of L has a negative eigenvalue. Indeed, let $\alpha \notin S$ and $a, b \in \mathbb{R}^H$. The conditions above give

$$\begin{aligned} & L(A + tav_\alpha^t, B + tu_\alpha b^t) \\ &= L(A, B) + t^2 \left(\frac{1}{2} |Ba|^2 + \frac{1}{2} |A^t b|^2 - s_\alpha \langle a, b \rangle \right) + \frac{t^4}{2} \langle a, b \rangle^2 \quad \text{for } t \in \mathbb{R}. \end{aligned}$$

Since $\text{rank } P = |S| < r \leq H$, Sylvester's rank inequality

$$\text{rank } P \geq \text{rank } A + \text{rank } B - H$$

shows that $\ker B$ or $\ker A^t$ is non-zero. If $Ba = 0$ with $a \neq 0$, then $b = \delta a$ with a small $\delta > 0$ makes the coefficient of t^2 negative. If $A^t b = 0$ with $b \neq 0$, take $a = \delta b$. If a unit vector a lies in both kernels, then $b = a$ gives

$$L(A + tav_\alpha^t, B + tu_\alpha a^t) = L(A, B) - s_\alpha t^2 + \frac{t^4}{2}.$$

Its minimum, at $t^2 = s_\alpha$, adds the mode α to BA . This is the direction along which gradient flow leaves C_k in Theorem 3.3 below, with $\alpha = k + 1$ and $a = e_{k+1}$.

The following theorem describes gradient flow from a small aligned initialization. To simplify the formulas, we take the learning rate in gradient flow to be 1; a time constant τ in front of $\dot{A}$ and $\dot{B}$ only rescales time.

Theorem 3.3 (Saxe-McClelland-Ganguli). *Let $e_1, \dots, e_H$ be the standard basis of $\mathbb{R}^H$ and $a_0 > 0$. Consider gradient flow $\dot{A} = -\partial_A L$, $\dot{B} = -\partial_B L$ for (3.1) with the initial condition*

$$A(0) = \sqrt{a_0} \sum_{\alpha=1}^r e_\alpha v_\alpha^t \quad \text{and} \quad B(0) = \sqrt{a_0} \sum_{\alpha=1}^r u_\alpha e_\alpha^t.$$

Then for all $t \geq 0$,

$$A(t) = \sum_{\alpha=1}^r \sqrt{a_\alpha(t)} e_\alpha v_\alpha^t, \quad B(t) = \sum_{\alpha=1}^r \sqrt{a_\alpha(t)} u_\alpha e_\alpha^t, \quad \text{and} \quad BA = \sum_{\alpha=1}^r a_\alpha(t) u_\alpha v_\alpha^t,$$

where each mode strength a_α solves the logistic equation $\dot{a}_\alpha = 2a_\alpha(s_\alpha - a_\alpha)$ with $a_\alpha(0) = a_0$, i.e.,

$$a_\alpha(t) = \frac{s_\alpha}{1 + \left(\frac{s_\alpha}{a_0} - 1 \right) e^{-2s_\alpha t}}.$$

In particular, if $a_0 < \frac{s_\alpha}{2}$, then $a_\alpha(t) = \frac{s_\alpha}{2}$ at the time

$$t_\alpha = \frac{1}{2s_\alpha} \log \left(\frac{s_\alpha}{a_0} - 1 \right) = \frac{1}{2s_\alpha} \log \frac{s_\alpha}{a_0} + O \left(\frac{a_0}{s_\alpha^2} \right).$$

This theorem is due to Saxe-McClelland-Ganguliⁱ.

ⁱA. Saxe, J. McClelland, and S. Ganguli, *Exact solutions to the nonlinear dynamics of learning in deep linear neural networks*. International Conference on Learning Representations, 2014.

PROOF. We look for a solution of the form $A = \sum_{\alpha} c_{\alpha} e_{\alpha} v_{\alpha}^t$ and $B = \sum_{\alpha} d_{\alpha} u_{\alpha} e_{\alpha}^t$ with real functions c_{α}, d_{α} . Then $BA = \sum_{\alpha} c_{\alpha} d_{\alpha} u_{\alpha} v_{\alpha}^t$ by orthonormality, and

$$BA - \Phi = \sum_{\alpha=1}^r (c_{\alpha} d_{\alpha} - s_{\alpha}) u_{\alpha} v_{\alpha}^t.$$

Since $\partial_A L = B^t(BA - \Phi)$ and $\partial_B L = (BA - \Phi)A^t$, we get

$$\partial_A L = \sum_{\alpha=1}^r d_{\alpha} (c_{\alpha} d_{\alpha} - s_{\alpha}) e_{\alpha} v_{\alpha}^t \quad \text{and} \quad \partial_B L = \sum_{\alpha=1}^r c_{\alpha} (c_{\alpha} d_{\alpha} - s_{\alpha}) u_{\alpha} e_{\alpha}^t.$$

So gradient flow preserves the form of (A, B) , and reduces to the decoupled system

$$\dot{c}_{\alpha} = -d_{\alpha} (c_{\alpha} d_{\alpha} - s_{\alpha}) \quad \text{and} \quad \dot{d}_{\alpha} = -c_{\alpha} (c_{\alpha} d_{\alpha} - s_{\alpha}) \quad \text{for } \alpha = 1, \dots, r.$$

By uniqueness of solutions of ordinary differential equations, this gives the solution. Moreover,

$$\frac{d}{dt} (c_{\alpha}^2 - d_{\alpha}^2) = 2c_{\alpha} \dot{c}_{\alpha} - 2d_{\alpha} \dot{d}_{\alpha} = 0.$$

So the balance $c_{\alpha} = d_{\alpha}$ at $t = 0$ is preserved. Then $a_{\alpha} = c_{\alpha} d_{\alpha} = c_{\alpha}^2$ satisfies

$$\dot{a}_{\alpha} = \dot{c}_{\alpha} d_{\alpha} + c_{\alpha} \dot{d}_{\alpha} = - (c_{\alpha}^2 + d_{\alpha}^2) (a_{\alpha} - s_{\alpha}) = 2a_{\alpha} (s_{\alpha} - a_{\alpha}).$$

The logistic equation is solved by separation of variables, which gives the formula for $a_{\alpha}(t)$. Finally, $a_{\alpha}(t) = \frac{s_{\alpha}}{2}$ if and only if

$$\left(\frac{s_{\alpha}}{a_0} - 1 \right) e^{-2s_{\alpha}t} = 1.$$

□

Corollary 3.4 (The staircase). *Along the flow in Theorem 3.3,*

$$L(A(t), B(t)) = \frac{1}{2} \sum_{\alpha=1}^r (s_{\alpha} - a_{\alpha}(t))^2.$$

As $a_0 \rightarrow 0$, the switching times satisfy $t_1 < t_2 < \dots < t_r$ and

$$t_{\alpha} \sim \frac{1}{2s_{\alpha}} \log \frac{1}{a_0}.$$

Each mode switches from $a_{\alpha} \approx 0$ to $a_{\alpha} \approx s_{\alpha}$ within a time window of order $\frac{1}{s_{\alpha}}$ around t_{α} . Hence, for $t_k \ll t \ll t_{k+1}$, the loss is close to L_k in (3.2), and the trajectory is close to the saddle set C_k .

PROOF. The formula for the loss follows from

$$\left\| \sum_{\alpha} c_{\alpha} u_{\alpha} v_{\alpha}^t \right\|_F^2 = \sum_{\alpha} c_{\alpha}^2.$$

The other claims follow from the explicit solution in Theorem 3.3, since $s_1 > s_2 > \dots > s_r$. □

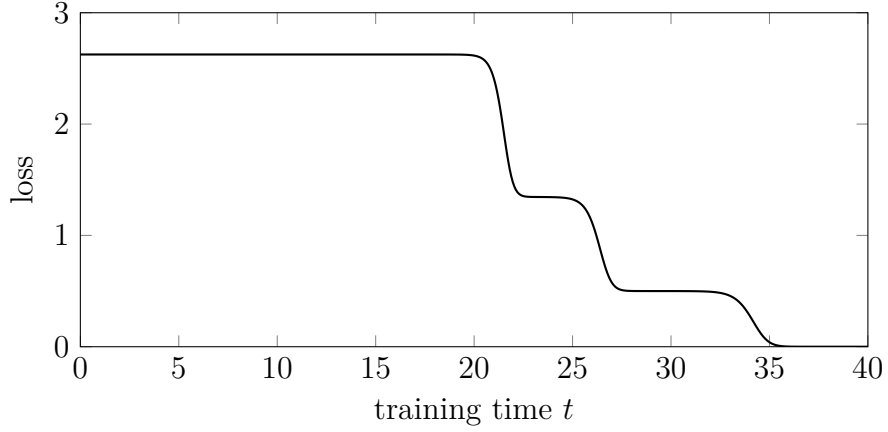


FIGURE 3.2. The staircase in Corollary 3.4 for $(s_1, s_2, s_3) = (1.6, 1.3, 1)$ and $a_0 = 10^{-30}$

So the loss falls in a staircase; see Figure 3.2. Long plateaus near the saddle sets $C_0, C_1, \dots$ are separated by sharp drops, as the modes are learned one at a time in the order of decreasing singular values. In this model, the question “which structure does the network learn, and when?” has a closed-form answer. We return to it in Chapter 7.

Part 2

AI safety

CHAPTER 4

What is AI safety?

AI safety is the study of how to build and deploy AI systems whose behavior is beneficial to humanity, and how to prevent the harms that powerful AI may cause. It draws on computer science, statistics, philosophy, economics, and, increasingly, mathematics. Its questions are no longer hypothetical: chat systems built on GPT-4, GPT-5, and GPT-6 are used every day by a very large number of people, and they increasingly act on their own. In this chapter, we describe the concerns of AI safety and two directions of research, interpretability and generalization out of distribution, with an eye toward the questions that become mathematics in Part III. Our presentation mainly follows Levine [L].

4.1. Risks from powerful AI

The main concerns of AI safety are the following.

- (1) *Misuse*: People use AI systems to cause harm on purpose.
- (2) *Misalignment*: An AI system pursues a goal different from the intended one, either because the goal was specified incorrectly or because training instilled a different goal (Sections 4.2 and 4.4).
- (3) *A lack of robustness*: An AI system which performs well on its training and test data fails on inputs unlike them (Section 4.4).
- (4) *A lack of legibility and steerability*: Humans cannot understand how an AI system computes, so they cannot verify or monitor its reasoning (Section 4.3), and they cannot reliably correct its behavior.
- (5) *A failure of cooperation*: Interacting AI systems and humans produce bad collective outcomes, even if each system behaves well alone.
- (6) *The gradual disempowerment of humans*: Humans delegate more and more decisions to systems that they do not understand, and they lose their influence over the institutions that shape their lives.

In these notes, we focus on one aspect, which lies between misalignment, robustness, and legibility: among the many parameters that fit the training data, which one does training select, and when? We explain in Section 4.2 and Chapter 5 why this is a safety question.

4.2. Where things go wrong

An AI system is produced in five stages:

- (1) *define* the desired behavior,
- (2) *measure* it by a score,
- (3) *train* the system on data,
- (4) *test* it on held-out data,
- (5) *deploy* it.

Each stage has its own failures: the definition may miss what was intended, the score may reward the wrong thing, the training data may omit important cases, and the test may share the

blind spots of the training data. For example, suppose that a content recommender is intended to benefit its users, but its success is measured by the time users spend on the platform. Then content that leaves users angry or addicted may score well merely because it holds their attention.

The training stage is where the mathematics of Part III enters. A trained network is one of many parameters that fit the training data, and these parameters may behave differently after deployment. So the question which parameter training selects is a safety question. We develop it in Chapter 5.

4.3. Interpretability

An AI system is *legible* to the extent that humans can understand how it represents concepts internally. Current training methods produce illegible AI: the engineers who design an AI system often cannot understand how it computes. The research field which tries to make AI more legible is called *interpretability*. It tries to reverse-engineer a trained network into human-understandable structure, validated by interventions on the network and not only by human-readable descriptions; see Olah et al.ⁱ. Its basic unit is a *feature*: a variable or property that the network represents in its activations, such as “the text is in French”.

Example (Handwritten digits). For the digit network, a feature may be “the image contains a closed loop”, which is present in the digits 0, 6, 8, and 9. Interpretability asks whether some neuron, or some direction in the 16-dimensional activations of a hidden layer, represents this feature in the trained network. It also asks whether changing the activations along this direction changes the predictions as expected.

A network with a few thousand dimensions per layer seems to represent millions of features. The *superposition hypothesis* of Elhage et al. [EHO] explains this as follows: when features are sparse, i.e., only a few are active at a time, a network can store more features than dimensions, as directions in the activation space that are not orthogonal but only nearly so. We state this hypothesis precisely in Chapter 11.

Example (Five features in the plane). Let $f_1, \dots, f_5 \in \mathbb{R}^2$ be the unit vectors at the angles $\frac{2\pi k}{5}$ for $k = 0, \dots, 4$, i.e., the vertices of a regular pentagon. No two of them are orthogonal. If features 1 and 2 are active, the activation is

$$y = f_1 + f_2,$$

and reading off feature 1 by projection gives

$$\langle y, f_1 \rangle = 1 + \cos \frac{2\pi}{5} \approx 1.31$$

instead of 1. The excess is the interference from feature 2. It is tolerable as long as few features are active at once.

ⁱC. Olah, N. Cammarata, L. Schubert, et al., *Zoom in: an introduction to circuits*. Distill 5 (2020).

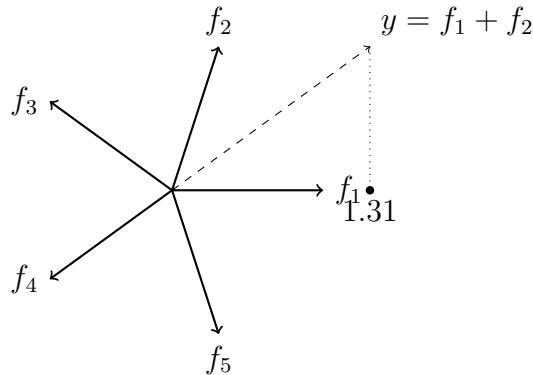


FIGURE 4.1. Five features in the plane; the projection of $f_1 + f_2$ onto f_1 is 1.31 instead of 1

The tool used in practice to read features out of a network is the *sparse autoencoder*, which represents the activations as sparse combinations of learned directions. The question when such sparse combinations can be decoded is a subject of Chapter 11.

A second success of interpretability is the reverse-engineering of learned algorithms. Networks trained on modular addition learn an algorithm based on the discrete Fourier transform on $\mathbb{Z}/p\mathbb{Z}$ [NCL]. Networks trained on multiplication in a finite group G learn algorithms based on the irreducible representations of G [CCN]. Which algorithm, and which representations, a given training run selects is not predicted by any theorem. We return to these problems in Chapter 12.

4.4. Generalization in and out of distribution

Two networks can perform equally well on the training data and yet compute their outputs in essentially different ways, and hence generalize differently. For this reason, testing and evaluation alone do not give assurances of safety, as argued by Lehallieur et al.ⁱ. The safety-critical regime is out of distribution: how a trained system behaves on inputs unlike those seen in training.

Example (Goal misgeneralization). In a video game, an agent was trained by reinforcement learning to collect a coin. In every training level the coin sat at the right end of the level. When the experimenters moved the coin, the agent ran past it to the right end of the level. It had learned “move right”, not “get the coin”. The two policies agree on every training level, so the data could not distinguish them. Yet training reliably produced one and not the other, as observed by Langosco et al.ⁱⁱ.

The example shows the general pattern: the system keeps its capabilities but pursues a different goal from the intended one. Training selects among the policies that fit the data by something other than the data: the architecture, the input encoding, the initialization, and the dynamics of training.

ⁱS. Lehallieur, J. Hoogland, M. Farrugia-Roberts, et al., *You are what you eat – AI alignment requires understanding how data shapes structure and generalisation*. arXiv:2502.05475.

ⁱⁱL. Langosco, J. Koch, L. Sharkey, et al., *Goal misgeneralization in deep reinforcement learning*. Proc. Mach. Learn. Res. 162 (2022), 12004–12019.

CHAPTER 5

Which solution does training find?

In this chapter, we formulate the question that connects AI safety with the mathematics of Part III: among the many parameters that fit the training data, which one does training select, and when? We describe two accounts of the selection, one Bayesian and one dynamical, and the program of developmental interpretability which tries to connect them. As in Chapter 4, our presentation mainly follows Levine [L].

5.1. Many solutions fit the data

A large network has many settings of its weights that fit the training data perfectly, and they may disagree everywhere else. The optimizer must land on one of them. By Section 2.2, the set W_0 of optimal parameters is typically a positive-dimensional set with singularities, and different parts of W_0 may implement different internal mechanisms. A held-out test drawn from the training distribution cannot tell two mechanisms apart if they agree on that distribution; see Figure 5.1. So the selection among the points of W_0 is made by the training process itself. Understanding it is a prerequisite for knowing whether a system that behaves as intended on the tests will behave as intended elsewhere, as argued by Lehalleur et al. (Section 4.4).

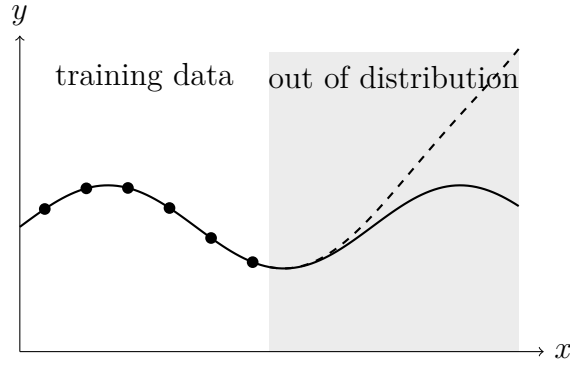


FIGURE 5.1. Two functions that agree on the training data (dots) and differ out of distribution (shaded)

Example (Selection by the initialization). Let $X \in \mathbb{M}_{n,d}(\mathbb{R})$ have rank $n < d$, and let $y \in \mathbb{R}^n$. Consider the linear model $f_w(x) = w \cdot x$ and the data points (x_j, y_j) , where $x_j^\dagger$ is the j -th row of X . With the square loss, the empirical loss is $L_n = \frac{2}{n}L$, where

$$L(w) = \frac{1}{2} \|Xw - y\|^2.$$

Its optimal set

$$W_0 = \{w : Xw = y\}$$

is an affine subspace of dimension $d - n$, and the data do not distinguish its points. Gradient flow of L is

$$\dot{w} = -X^t(Xw - y).$$

It moves only in the range of X^t , which is the orthogonal complement of $\ker X$. So

$$w(t) - w(0) \perp \ker X \quad \text{for } t \geq 0.$$

Since

$$\frac{d}{dt}(Xw(t) - y) = -XX^t(Xw(t) - y)$$

and XX^t is positive definite, $Xw(t) \rightarrow y$ exponentially fast. Hence $\dot{w}$ is integrable, and $w(t)$ converges to the unique point of W_0 in $w(0) + (\ker X)^\perp$, i.e., to the orthogonal projection of $w(0)$ onto W_0 . In particular, from $w(0) = 0$, gradient flow selects the solution of minimal norm,

$$X^t(XX^t)^{-1}y.$$

So the selection is made by the initialization and the dynamics. With noise, it is slowly forgotten. For the Langevin dynamics, L does not depend on the component of w_t in $\ker X$. So this component is its initial value plus $\sqrt{2\varepsilon}$ times a Brownian motion in $\ker X$, and the parameter moves a distance of order one along W_0 in a time of order ε^{-1} . The heuristic in Chapter 10 predicts the same time scale for the entropic drift. There is no stationary law, since $e^{-\frac{L}{\varepsilon}}$ is not integrable on $\mathbb{R}^d$.

5.2. Two accounts of selection

There are two theories that claim to predict the selection.

The first account is Bayesian. Singular learning theory, due to Watanabe [W], describes the Bayesian posterior distribution of the parameters given the data. It weighs regions of parameter space by volume, through an invariant of the singularities of L called the learning coefficient λ . Smaller λ means more nearly optimal parameters nearby, i.e., a flatter region, and the posterior prefers the regions of smallest λ . We develop this theory in Chapter 6.

The second account is dynamical. Training is not Bayesian; it is a dynamical process, gradient descent with noise. In the deep linear network, gradient flow selects structures one at a time, in the staircase of Corollary 3.4. We develop this account in Chapters 7, 8, and 10.

Whether the two accounts agree is a central question. The problem MAIS-O7 (Chapter 7) asks to build the bridge between them first in the deep linear network, where both sides are explicit.

Remark (Equilibrium and finite times). The two accounts agree at equilibrium. By Section 1.5 and the remark on the Langevin dynamics as a sampler in Section 3.3, the posterior is the stationary law of the Langevin dynamics with the potential $L_n - \frac{1}{n} \log \varphi$ at the temperature $\varepsilon = \frac{1}{n}$.

The two accounts differ at finite times. For the Langevin dynamics w_t with a potential L on $\mathbb{T}^d$ at temperature ε , the law of w_t approaches the normalized Gibbs measure on the time scale γ_ε^{-1} , where γ_ε is the spectral gap of the generator; see the remark on spectral gap and relaxation in Section 8.6. Suppose that L has exactly two local minima. Let h be the communication height between them (Chapter 8) minus the value of L at the shallower one. By the theorem of Holley-Kusuoka-Stroock in Chapter 10,

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \gamma_\varepsilon = -h.$$

Under the assumptions of Section 8.6, γ_ε^{-1} is, up to a factor $1 + o(1)$, the Eyring-Kramers time of the transition from the shallower well to the deeper one. For $t \ll \gamma_\varepsilon^{-1}$, the law of w_t is still concentrated in the wells reached from the initialization, with the weights inherited from the law of w_0 and not with the Gibbs weights. Moreover, in training, the temperature is set by the learning rate and the minibatch size (Section 3.3), not by n , and the noise is not isotropic. In this sense, Bayesian selection is the limit $t \rightarrow \infty$, while training stops at a finite time.

5.3. Stages of learning and developmental interpretability

The learning coefficient can be estimated at the parameter that a training run has reached, by sampling with the Langevin dynamics. This local estimate measures how degenerate the current solution is [LFW]. As a network trains, its estimated local learning coefficient is observed to jump at discrete moments, called *phase transitions*. At such a moment, the solution moves to a qualitatively new, differently degenerate region of parameter space, as a new piece of internal structure forms. Tracking these jumps is the program of *developmental interpretability*.

For safety, the hope is to predict the stages of a training run, and hence the arrival of new capabilities, from geometric quantities measurable along the way. In stagewise training, new capabilities may arrive abruptly, as in the grokking of Section 3.4.

Remark (Phase transitions). In statistical mechanics, a phase transition is a value of a parameter at which the equilibrium state changes abruptly and the free energy is not smooth. In Bayesian learning, the parameter is n . Let U_1 and U_2 be disjoint neighborhoods of two sets of local minimizers with the losses $L_1 > L_2$ and the learning coefficients $\lambda_1 < \lambda_2$. With the empirical entropy S_n of Chapter 6, we expect, as in (7.1),

$$F_n(U_i) \approx nS_n + nL_i + \lambda_i \log n \quad \text{for } i = 1, 2.$$

Since $\pi_n(U) = e^{F_n - F_n(U)}$ (Chapter 6),

$$\frac{\pi_n(U_1)}{\pi_n(U_2)} = e^{F_n(U_2) - F_n(U_1)} \approx e^{-n(L_1 - L_2) + (\lambda_2 - \lambda_1) \log n}.$$

So the posterior mass moves from U_1 to U_2 near the large root n_c of

$$\frac{n}{\log n} = \frac{\lambda_2 - \lambda_1}{L_1 - L_2},$$

as in (7.2). The exponent changes by $O(1)$ on a window of n of length $O(\frac{1}{L_1 - L_2})$, which is small compared to n_c when n_c is large. Moreover, since U_1 and U_2 are disjoint,

$$\min_i F_n(U_i) - \log 2 \leq F_n(U_1 \cup U_2) = -\log(e^{-F_n(U_1)} + e^{-F_n(U_2)}) \leq \min_i F_n(U_i).$$

So $F_n(U_1 \cup U_2)$ is, up to $\log 2$, the minimum of the two local free energies, whose leading terms cross at n_c with a corner. This is the Bayesian ladder of Section 7.1. The phase transitions of developmental interpretability occur in training time instead: the estimated local learning coefficient at w_t jumps when w_t leaves a plateau near one critical set for the next one. In Corollary 3.4, a plateau lasts a time of order $\log \frac{1}{a_0}$, while the jump takes a time of order $\frac{1}{s_\alpha}$. The conjecture MAIS-O7 in Chapter 7 asserts that, in the deep linear network, the learning coefficient rises as the loss falls along the saddle sets visited by gradient flow, as it does along the Bayesian ladder.

5.4. Two clocks

The Bayesian account and the dynamical account run on different clocks. The Bayesian clock ticks in the sample size n : as n grows, the posterior moves to regions of lower loss, trading loss against the learning coefficient. The dynamical clock ticks in training time t : gradient descent escapes a saddle point at a rate set by the curvature of the loss there.

Remark (The clock at a saddle point). Let z be a Morse saddle of L , i.e., a non-degenerate critical point of index one (Chapter 8), and let $-\mu < 0$ be the negative eigenvalue of $\nabla^2 L(z)$. Near z , gradient flow is approximately linear, and the component of $w(t) - z$ in the unstable direction grows like $e^{\mu t}$. So a trajectory which starts near z , at the distance δ from the stable manifold of z , leaves a fixed neighborhood of z after the time

$$\frac{1}{\mu} \log \frac{1}{\delta} + O(1) \quad \text{as } \delta \rightarrow 0.$$

In the proof of Theorem 3.3, the linearization of the system for (c_α, d_α) at the origin is

$$\dot{c}_\alpha = s_\alpha d_\alpha \quad \text{and} \quad \dot{d}_\alpha = s_\alpha c_\alpha,$$

with the eigenvalues $\pm s_\alpha$, and the stable manifold of the origin is the line $\{c_\alpha = -d_\alpha\}$. The initial point $(\sqrt{a_0}, \sqrt{a_0})$ is at the distance $\sqrt{2a_0}$ from it. So the formula above gives

$$t_\alpha = \frac{1}{2s_\alpha} \log \frac{1}{a_0} + O(1),$$

as in Theorem 3.3. With noise, the process leaves even from the saddle point itself. In dimension one, the Langevin dynamics of $L(x) = -\frac{\mu}{2}x^2$ with $x_0 = 0$ is

$$x_t = \sqrt{2\varepsilon} e^{\mu t} \int_0^t e^{-\mu s} dB_s,$$

and $e^{-\mu t}x_t$ converges almost surely to a Gaussian random variable with mean 0 and variance $\frac{\varepsilon}{\mu}$. So $|x_t|$ first reaches a fixed $\delta_0 > 0$ at the time

$$\frac{1}{2\mu} \log \frac{1}{\varepsilon} + O_p(1),$$

where $O_p(1)$ is a term bounded in probability. Hence, at a saddle point, the dynamical clock is logarithmic in the initial distance or in the temperature, with the rate μ set by the curvature. At a well, it is exponential, of order $e^{\frac{h}{\varepsilon}}$ (Chapter 8). The Bayesian clock has no analogue of μ : the coefficient of $\log n$ in the free energy of a window around a saddle point is decided by volumes and by the rim of the window (Section 7.3).

The two accounts also use different data. The selection of a region is decided by volumes, and it is measured by the learning coefficient. The time that training spends near a region is not: rescaling L by a constant changes no learning coefficient but changes every time scale. Relating the two clocks is a theme of Part III (Chapters 7 and 8).

Remark (Rescaling the loss). Let $c > 0$. By Lemma 6.6(1), cL has the same local invariants (λ, m) as L . Time scales are not invariant. If $w(t)$ is a gradient flow of L , then $w(ct)$ is a gradient flow of cL , and the Hessian eigenvalues, which set the escape rates from saddle points, are multiplied by c . For the Langevin dynamics, let w_t solve

$$dw_t = -c\nabla L(w_t) dt + \sqrt{2\varepsilon} dB_t.$$

Then $v_s = w_{\frac{s}{c}}$ solves

$$dv_s = -\nabla L(v_s) ds + \sqrt{\frac{2\varepsilon}{c}} d\tilde{B}_s,$$

where $\tilde{B}_s = \sqrt{c} B_{\frac{s}{c}}$ is again a standard Brownian motion. So the dynamics of cL at temperature ε is the dynamics of L at temperature $\frac{\varepsilon}{c}$, run c times faster. If $\tau^{L,\varepsilon}$ denotes the hitting time of a fixed set by the Langevin dynamics of L at temperature ε from a fixed initial point, then

$$\tau^{cL,\varepsilon} = \frac{1}{c} \tau^{L,\frac{\varepsilon}{c}} \quad \text{in law.}$$

In particular, the Arrhenius factor $e^{\frac{h}{\varepsilon}}$ becomes $e^{\frac{ch}{\varepsilon}}$. In the singular Eyring-Kramers law of Theorem 9.4, the learning coefficient λ_A of the starting well enters the dependence on ε only through the power $\varepsilon^{\lambda_A - \frac{d}{2}}$, which the rescaling does not change. The exponential rate is the barrier height h , which is not an invariant of the singularities of L .

5.5. From questions to problems

The MAIS repository of Levine [L], at github.com/Math-for-AI-Safety/MAIS, turns such questions into precisely stated open problems. The problems labeled MAIS-On in these notes are from its problem list. In Part III, we treat the following problems in detail:

- Chapter 6: MAIS-O69 on the estimation of the local learning coefficient, MAIS-O70 and MAIS-O77 on local learning coefficients of the deep linear network, and MAIS-O72 on the stratification of W_0 by the learning coefficient;
- Chapter 7: MAIS-O7 on opposing staircases, and MAIS-O79 and MAIS-O81 on saddles;
- Chapter 9: MAIS-O73 and MAIS-O74 on Eyring-Kramers laws for singular wells;
- Chapter 10: MAIS-O75 on entropic selection;
- Chapter 11: MAIS-O4, O40, O42, and O47 on superposition;
- Chapter 12: MAIS-O6, O59, O62, and O63 on modular addition.

We chose these problems because they are the closest to semiclassical analysis, spectral theory, and harmonic analysis. In Chapter 13, we suggest where to start.

Part 3

Mathematics of AI safety

CHAPTER 6

Singular learning theory

Singular learning theory, developed by Watanabe [W], is the asymptotic theory of Bayesian learning in singular models. Its central invariant, the learning coefficient, measures the volume of nearly optimal parameters, and it replaces the dimension count of classical statistics. In this chapter, we introduce Bayesian learning and the free energy, compute the free energy in the regular case, define the learning coefficient, and state Watanabe's theorem. We then discuss the local learning coefficient, its estimation, the stratification of the optimal set, and the computation of learning coefficients, which lead to the problems MAIS-O69, O70, O72, and O77. Our presentation mainly follows Watanabe [W] and, for the local learning coefficient, Lau et al. [LFW].

6.1. Bayesian learning and the free energy

Throughout this chapter, we consider a statistical model $p(x \mid w)$ without inputs, with parameters in a compact set $W \subset \mathbb{R}^d$. We assume that it is realizable, i.e., the data $X_1, \dots, X_n$ are independent samples from $q_0 = p(\cdot \mid w_0)$ for some $w_0 \in W$. By Proposition 1.1, the loss

$$L(w) = \text{KL}(q_0 \parallel p(\cdot \mid w))$$

is non-negative, and $W_0 = L^{-1}(0)$ is the set of parameters that reproduce the truth.

Definition (Prior, posterior, and free energy). A prior is a probability density φ on W , which encodes which parameters are plausible before any data are seen. The posterior is the probability measure

$$\pi_n(dw) = \frac{1}{Z_n} \varphi(w) \prod_{j=1}^n p(X_j \mid w) dw, \quad \text{where } Z_n = \int_W \varphi(w) \prod_{j=1}^n p(X_j \mid w) dw.$$

The free energy is

$$F_n = -\log Z_n.$$

For a measurable set $U \subset W$, the local free energy $F_n(U)$ is defined in the same way, with the integral restricted to U .

The number Z_n is the probability that the model, averaged over the prior, assigns to the sample. So F_n is small when the model predicts the data well. Regions of small local free energy carry most of the posterior mass, since

$$\pi_n(U) = e^{F_n - F_n(U)}.$$

Write

$$S_n = -\frac{1}{n} \sum_{j=1}^n \log q_0(X_j) \quad \text{and} \quad L_n(w) = \frac{1}{n} \sum_{j=1}^n \log \frac{q_0(X_j)}{p(X_j \mid w)}.$$

Then S_n is the empirical entropy, and L_n is the empirical loss, normalized so that $\mathbb{E}[L_n(w)] = L(w)$. Since

$$\prod_{j=1}^n p(X_j \mid w) = e^{-nS_n} e^{-nL_n(w)},$$

we obtain

$$F_n = nS_n - \log \int_W \varphi(w) e^{-nL_n(w)} dw.$$

The first term does not depend on the model. The second term is controlled by the deterministic integral

$$Z(n) = \int_W \varphi(w) e^{-nL(w)} dw,$$

in which L_n is replaced by its mean L . By Watanabe's theory [W], the replacement costs a random term which is bounded in probability; see the remark on the random term in Section 6.3. We therefore study the asymptotics of $Z(n)$ as $n \rightarrow \infty$. It is a Laplace integral, with the large parameter n and the phase L .

6.2. The regular case

Proposition 6.1. *Suppose that $L \in C^3(W)$, $W_0 = \{w_0\}$ with w_0 in the interior of W , the Hessian $I = \nabla^2 L(w_0)$ is positive definite, and φ is continuous and positive. Then*

$$Z(n) = \varphi(w_0) \left(\frac{2\pi}{n} \right)^{\frac{d}{2}} (\det I)^{-\frac{1}{2}} (1 + o(1)) \quad \text{as } n \rightarrow \infty.$$

Consequently,

$$-\log Z(n) = \frac{d}{2} \log n + O(1).$$

PROOF. Let $\theta \in (0, 1)$. By Taylor's formula and the continuity of φ , there is $\delta > 0$ such that for $|w - w_0| \leq \delta$,

$$(1 - \theta) \frac{1}{2} (w - w_0)^t I (w - w_0) \leq L(w) \leq (1 + \theta) \frac{1}{2} (w - w_0)^t I (w - w_0)$$

and

$$(1 - \theta) \varphi(w_0) \leq \varphi(w) \leq (1 + \theta) \varphi(w_0).$$

On the compact set $\{w \in W : |w - w_0| \geq \delta\}$, we have $L \geq c$ for some $c > 0$. So this set contributes at most

$$e^{-nc} \max \varphi \cdot \text{Vol}(W) = o\left(n^{-\frac{d}{2}}\right)$$

to $Z(n)$. Write

$$Z_0(n) = \varphi(w_0) \left(\frac{2\pi}{n} \right)^{\frac{d}{2}} (\det I)^{-\frac{1}{2}}.$$

On the ball, the substitution $w - w_0 = n^{-\frac{1}{2}} v$ and the Gaussian integral

$$\int_{\mathbb{R}^d} e^{-\frac{1 \pm \theta}{2} v^t I v} dv = \left(\frac{2\pi}{1 \pm \theta} \right)^{\frac{d}{2}} (\det I)^{-\frac{1}{2}}$$

show that

$$(1 - \theta)(1 + \theta)^{-\frac{d}{2}} \leq \liminf_{n \rightarrow \infty} \frac{Z(n)}{Z_0(n)} \leq \limsup_{n \rightarrow \infty} \frac{Z(n)}{Z_0(n)} \leq (1 + \theta)(1 - \theta)^{-\frac{d}{2}}.$$

For the lower bound, we restrict the Gaussian integral to $|v| \leq \delta\sqrt{n}$, which loses $o(1)$. Letting $\theta \rightarrow 0$ completes the proof. $\square$

So in a regular model,

$$F_n = nS_n + \frac{d}{2} \log n + O_p(1),$$

where $O_p(1)$ denotes a term bounded in probability. This is the expansion behind the Bayesian information criterion of Schwarzⁱ. The term $\frac{d}{2} \log n$ is the price of not knowing w_0 : the posterior concentrates in a ball of radius of order $n^{-\frac{1}{2}}$, whose prior volume is of order $n^{-\frac{d}{2}}$. Every parameter costs half a $\log n$. This is the classical complexity penalty, which compares models by their numbers of parameters.

6.3. The learning coefficient

In a singular model, W_0 is not a point and L is not quadratic near W_0 , so Proposition 6.1 does not apply. The right replacement is a volume law.

Definition (Volume function). The volume function of L with respect to the prior φ is

$$V(\varepsilon) = \int_{\{w \in W: L(w) < \varepsilon\}} \varphi(w) dw \quad \text{for } \varepsilon > 0.$$

Proposition 6.2. *For every $n > 0$,*

$$Z(n) = \int_0^\infty e^{-u} V\left(\frac{u}{n}\right) du.$$

If, moreover,

$$V(\varepsilon) = c\varepsilon^\lambda \left(\log \frac{1}{\varepsilon}\right)^{m-1} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0^+$$

for some $c, \lambda > 0$ and an integer $m \geq 1$, then

$$Z(n) = \Gamma(\lambda + 1) c n^{-\lambda} (\log n)^{m-1} (1 + o(1)) \quad \text{as } n \rightarrow \infty.$$

In particular,

$$-\log Z(n) = \lambda \log n - (m - 1) \log \log n + O(1).$$

PROOF. Since $L \geq 0$,

$$e^{-nL(w)} = \int_0^\infty e^{-u} \mathbf{1}_{\{nL(w) < u\}} du,$$

and the first claim follows from Fubini's theorem. For the second claim, write

$$\varepsilon = \frac{1}{n} \quad \text{and} \quad N(\varepsilon) = c\varepsilon^\lambda \left(\log \frac{1}{\varepsilon}\right)^{m-1}.$$

Fix $\varepsilon_0 \in (0, e^{-1})$ such that

$$V(\varepsilon) \leq 2c\varepsilon^\lambda \left(\log \frac{1}{\varepsilon}\right)^{m-1} \quad \text{for } 0 < \varepsilon \leq \varepsilon_0.$$

Since $V \leq 1$,

$$\int_{\frac{\varepsilon_0}{\varepsilon}}^\infty e^{-u} V(\varepsilon u) du \leq e^{-\frac{\varepsilon_0}{\varepsilon}} = o(N(\varepsilon)).$$

For $0 < u < \frac{\varepsilon_0}{\varepsilon}$ and $\varepsilon < e^{-1}$, we have

$$\log \frac{1}{\varepsilon u} \leq \log \frac{1}{\varepsilon} + |\log u| \leq \log \frac{1}{\varepsilon} (1 + |\log u|).$$

ⁱG. Schwarz, *Estimating the dimension of a model*. Ann. Statist. 6 (1978), 461–464.

So

$$\frac{V(\varepsilon u)}{N(\varepsilon)} \leq 2u^\lambda (1 + |\log u|)^{m-1},$$

which is integrable against $e^{-u} du$ on $(0, \infty)$ since $\lambda > 0$. For each fixed $u > 0$,

$$\frac{V(\varepsilon u)}{N(\varepsilon)} = \frac{V(\varepsilon u)}{c(\varepsilon u)^\lambda (\log \frac{1}{\varepsilon u})^{m-1}} \cdot u^\lambda \left(\frac{\log \frac{1}{\varepsilon} - \log u}{\log \frac{1}{\varepsilon}} \right)^{m-1} \rightarrow u^\lambda.$$

By the dominated convergence theorem,

$$\frac{1}{N(\varepsilon)} \int_0^{\frac{\varepsilon_0}{\varepsilon}} e^{-u} V(\varepsilon u) du \rightarrow \int_0^\infty e^{-u} u^\lambda du = \Gamma(\lambda + 1).$$

This completes the proof. $\square$

Example.

- (1) [Regular case] In the setting of Proposition 6.1, $\{L < \varepsilon\}$ is approximately the ellipsoid

$$\left\{ \frac{1}{2} (w - w_0)^t I (w - w_0) < \varepsilon \right\}.$$

So

$$V(\varepsilon) = \varphi(w_0) \omega_d (2\varepsilon)^{\frac{d}{2}} (\det I)^{-\frac{1}{2}} (1 + o(1)),$$

where

$$\omega_d = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2} + 1\right)}$$

is the volume of the unit ball. So $\lambda = \frac{d}{2}$ and $m = 1$, and Proposition 6.2 recovers Proposition 6.1.

- (2) [A degenerate minimum] Let $W = [-1, 1]$, $\varphi = \frac{1}{2}$, and $L(w) = w^{2k}$. Then $V(\varepsilon) = \varepsilon^{\frac{1}{2k}}$ for $\varepsilon \leq 1$, so $\lambda = \frac{1}{2k}$ and $m = 1$.
- (3) [A singular optimal set] Let $W = [-1, 1]^2$, $\varphi = \frac{1}{4}$, and $L(a, b) = a^2 b^2$, as in the smallest singular network of Section 2.2. For $0 < s < 1$,

$$\text{Leb} \left(\left\{ (a, b) \in [-1, 1]^2 : |ab| < s \right\} \right) = 4 \int_0^1 \min \left\{ 1, \frac{s}{a} \right\} da = 4s \left(1 + \log \frac{1}{s} \right),$$

so with $s = \sqrt{\varepsilon}$,

$$V(\varepsilon) = \sqrt{\varepsilon} \left(1 + \frac{1}{2} \log \frac{1}{\varepsilon} \right) = \frac{1}{2} \varepsilon^{\frac{1}{2}} \log \frac{1}{\varepsilon} (1 + o(1)).$$

So $\lambda = \frac{1}{2}$ and $m = 2$: the exponent stays $\frac{1}{2}$ instead of rising to $\frac{d}{2} = 1$, and a logarithm appears; see Figure 6.1.

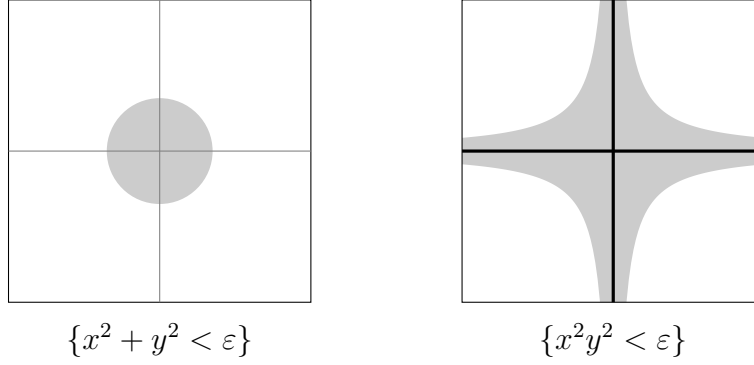


FIGURE 6.1. Nearly optimal parameters near a regular minimum (left) and near the singular optimal set $\{xy = 0\}$ (right); the second region is much larger for small ε

The following theorem says that the volume law always holds for analytic models; see Watanabe [W]. It rests on the resolution of singularities by Hironakaⁱ.

Theorem (Volume asymptotics). *Suppose that L is real-analytic and φ is smooth and positive on W . Then there are a rational number $\lambda > 0$, an integer $m \in \{1, \dots, d\}$, and a constant $c > 0$ such that*

$$V(\varepsilon) = c\varepsilon^\lambda \left(\log \frac{1}{\varepsilon} \right)^{m-1} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0^+.$$

Definition (Learning coefficient). The numbers λ and m are called the learning coefficient and the multiplicity of the model. In algebraic geometry, λ is known as the real log canonical threshold of L with respect to φ .

Remark (The zeta function). The exponents can be read off from the zeta function

$$\zeta(z) = \int_W L(w)^z \varphi(w) dw = \int_0^\infty \varepsilon^z dV(\varepsilon) \quad \text{for } \operatorname{Re} z > 0.$$

Since

$$\int_0^1 \varepsilon^{z+\lambda-1} \left(\log \frac{1}{\varepsilon} \right)^{m-1} d\varepsilon = \frac{\Gamma(m)}{(z+\lambda)^m},$$

a volume law with exponents (λ, m) corresponds to a pole of order m at $z = -\lambda$. For instance, the degenerate minimum w^{2k} above has

$$\zeta(z) = \frac{1}{2kz + 1},$$

and the singular optimal set $\{a^2 b^2 = 0\}$ above has

$$\zeta(z) = \frac{1}{(2z + 1)^2}.$$

By resolution of singularities, ζ extends meromorphically to $\mathbb{C}$ with poles at negative rational numbers, as shown by Atiyahⁱⁱ and Bernstein-Gelfandⁱⁱⁱ. The largest pole is $-\lambda$, and its order is m [W].

ⁱH. Hironaka, *Resolution of singularities of an algebraic variety over a field of characteristic zero. I, II*. Ann. of Math. (2) 79 (1964), 109–203, 205–326.

ⁱⁱM. Atiyah, *Resolution of singularities and division of distributions*. Comm. Pure Appl. Math. 23 (1970), 145–150.

ⁱⁱⁱI. Bernstein and S. Gelfand, *Meromorphic property of the functions P^λ* . Funct. Anal. Appl. 3 (1969), 68–69.

Concretely, resolution of singularities gives a proper real-analytic map g onto a neighborhood of W_0 , and finitely many charts covering its domain, with local coordinates u in which

$$L(g(u)) = u_1^{2k_1} \cdots u_d^{2k_d} \quad \text{and} \quad |\det g'(u)| = b(u) |u_1^{h_1} \cdots u_d^{h_d}|,$$

where $k_j, h_j \geq 0$ are integers, and b is real-analytic and positive. Since b is positive, the largest pole of the contribution of each chart to ζ , and its order, are those of the product of the one-variable integrals

$$\int_0^1 u^{2k_j z + h_j} du = \frac{1}{2k_j z + h_j + 1}.$$

Hence λ is the smallest of the numbers $\frac{h_j+1}{2k_j}$ over all charts and all j with $k_j > 0$, and m is the largest number of indices j attaining it in one chart. This is why λ is rational. For $L = a^2 b^2$, no resolution is needed: $k = (1, 1)$, $h = (0, 0)$, and $\frac{h_1+1}{2k_1} = \frac{h_2+1}{2k_2} = \frac{1}{2}$ give $\lambda = \frac{1}{2}$ and $m = 2$.

In general, $\zeta(-s) = \int_W L^{-s} \varphi dw$ is finite for $0 < s < \lambda$ and infinite for $s = \lambda$, by the volume law. Hence

$$\lambda = \sup \left\{ s > 0 : \int_W L(w)^{-s} \varphi(w) dw < \infty \right\}.$$

For $L(w) = |w|^2$, this is the fact that $|w|^{-2s}$ is integrable near 0 in $\mathbb{R}^d$ if and only if $s < \frac{d}{2}$. It also explains the name real log canonical threshold. In complex algebraic geometry, the log canonical threshold of a holomorphic function f is the supremum of the numbers $c > 0$ for which $|f|^{-2c}$ is locally integrable. For $L = f^2$ with real f , λ is the same supremum for real integrals.

Remark (Blow-ups as polar coordinates). The simplest blow-up is the passage to polar coordinates. Suppose that L is a homogeneous polynomial of degree $2k$ on $\mathbb{R}^d$, and let φ be the indicator function of the unit ball. Then

$$\zeta(z) = \int_0^1 r^{2kz+d-1} dr \int_{\mathbb{S}^{d-1}} L(\theta)^z d\sigma(\theta) = \frac{\zeta_{\mathbb{S}}(z)}{2kz+d},$$

where $\zeta_{\mathbb{S}}$ is the zeta function of L on the unit sphere with its surface measure σ . The radial factor has a simple pole at $-\frac{d}{2k}$, and the poles of $\zeta_{\mathbb{S}}$ come from the zeros of L on the sphere. Near a point $w \neq 0$ of W_0 , the factor r^{2k} is bounded above and below, so the local invariants $(\lambda(w), m(w))$ of L at w , defined in the next section, are those of $L|_{\mathbb{S}^{d-1}}$ at $\frac{w}{|w|}$. Hence $\lambda(0)$ is the smaller of $\frac{d}{2k}$ and the smallest local learning coefficient of $L|_{\mathbb{S}^{d-1}}$. If the two are equal, the orders of the poles add, and $m(0)$ is one more than the multiplicity on the sphere. If $L > 0$ on $\mathbb{S}^{d-1}$, then $\lambda(0) = \frac{d}{2k}$ and $m(0) = 1$, as for a positive definite quadratic form and for w^{2k} with $d = 1$. For $L = a^2 b^2$, we have $\frac{d}{2k} = \frac{1}{2}$, and the zeros of $\cos^2 \theta \sin^2 \theta$ on the circle are quadratic, so they also have $\lambda = \frac{1}{2}$. This gives $(\lambda, m) = (\frac{1}{2}, 2)$ at the origin.

In an affine chart, polar coordinates are the directional blow-up

$$w_1 = u_1 \quad \text{and} \quad w_j = u_1 u_j \quad \text{for } 2 \leq j \leq d,$$

for which

$$dw = |u_1|^{d-1} du \quad \text{and} \quad L(w) = u_1^{2k} L(1, u_2, \dots, u_d).$$

This is how the exponents h_j of the Jacobian arise in the normal crossing form: here $h_1 = d - 1$ and $k_1 = k$, so $\frac{h_1+1}{2k_1} = \frac{d}{2k}$. If $L(1, u_2, \dots, u_d)$ still vanishes, then it is resolved in $d - 1$ variables. The resolution theorem says that finitely many blow-ups, along points and more general smooth submanifolds, reach the normal crossing form. For $L = a^2 b^2$, the chart $a = u_1$, $b = u_1 u_2$ gives

$$L = u_1^4 u_2^2 \quad \text{and} \quad da db = |u_1| du.$$

So $k = (2, 1)$ and $h = (1, 0)$, and both ratios $\frac{h_j+1}{2k_j}$ equal $\frac{1}{2}$, in accordance with $m = 2$.

The loss $L(A, B) = \frac{1}{2}\|BA\|_F^2$ of the two-layer linear network with $\Phi = 0$ is a homogeneous polynomial of degree 4 in $d = H(M+N)$ variables. So its local invariants on F_0 are constant along rays, and $\lambda(0) \leq \frac{H(M+N)}{4}$. For $M = N = H = 2$, this bound is 2, while the Aoyagi-Watanabe coefficient of Section 6.6 is $\frac{3}{2}$. Since $\lambda(0)$ is the smaller of 2 and the values on the sphere, the value $\frac{3}{2}$ is attained at points $w \neq 0$ of F_0 .

Remark (Oscillatory integrals). The same exponents govern oscillatory integrals with the phase L , which are central in harmonic analysis. Suppose that L is real-analytic, that $\varphi \geq 0$ is smooth with compact support in the interior of W and positive somewhere on W_0 , and that L has no critical point in $\text{supp } \varphi \setminus W_0$. Then V , defined with this φ , satisfies a volume law with some constants $c > 0$, λ , and m . Moreover, V is smooth on $(0, \infty)$, and, by resolution of singularities, its asymptotic expansion as $\varepsilon \rightarrow 0^+$ can be differentiated term by term. Since

$$\int_W e^{i\tau L(w)} \varphi(w) dw = \int_0^\infty e^{i\tau\varepsilon} V'(\varepsilon) d\varepsilon,$$

the oscillatory analogue of Proposition 6.2 gives

$$\int_W e^{i\tau L(w)} \varphi(w) dw = \Gamma(\lambda + 1) c e^{\frac{i\pi\lambda}{2}} \tau^{-\lambda} (\log \tau)^{m-1} (1 + o(1)) \quad \text{as } \tau \rightarrow \infty;$$

see Arnold-Gusein-Zade-Varchenkoⁱ. So the learning coefficient is the decay rate of the oscillatory integral with the phase L , and the volume law is a sublevel set estimate for L ; see Phong-Stein-Sturmⁱⁱ and Steinⁱⁱⁱ.

The Newton polyhedron of L at a point of W_0 is the convex hull of the union of the orthants $\alpha + [0, \infty)^d$ over the exponents α of the Taylor series of L at that point. The Newton distance is the number t such that $(t, \dots, t)$ lies on its boundary. For the monomial $\prod_j w_j^{2k_j}$, the Newton polyhedron is the single orthant $(2k_1, \dots, 2k_d) + [0, \infty)^d$. So $t = \max_j 2k_j$, and $\lambda = \frac{1}{t}$ by the remark on the zeta function. Varchenko^{iv} showed that the decay rate $\tau^{-\frac{1}{t}}$, up to a power of $\log \tau$, persists for phases which are non-degenerate with respect to their Newton polyhedra and have Newton distance $t > 1$. The monomial wells of Chapter 9 are of this kind when G is real-analytic.

By Proposition 6.2,

$$-\log Z(n) = \lambda \log n - (m - 1) \log \log n + O(1).$$

Watanabe's theorem is the corresponding statement for the random free energy.

Theorem 6.3 (Watanabe). *For a singular model satisfying Watanabe's regularity conditions,*

$$F_n = nS_n + \lambda \log n - (m - 1) \log \log n + O_p(1) \quad \text{as } n \rightarrow \infty.$$

ⁱV. Arnold, S. Gusein-Zade, and A. Varchenko, *Singularities of Differentiable Maps, Volume II: Monodromy and Asymptotics of Integrals*. Birkhäuser, 1988.

ⁱⁱD. Phong, E. Stein, and J. Sturm, *On the growth and stability of real-analytic functions*. Amer. J. Math. 121 (1999), 519–554.

ⁱⁱⁱE. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*. Princeton University Press, 1993.

^{iv}A. Varchenko, *Newton polyhedra and estimation of oscillating integrals*. Funct. Anal. Appl. 10 (1976), 175–196.

Moreover, the expected generalization error, i.e., the expected Kullback-Leibler divergence of the Bayesian predictive distribution p_n from q_0 , satisfies

$$\mathbb{E} [\text{KL}(q_0 \| p_n)] = \frac{\lambda}{n} + o\left(\frac{1}{n}\right).$$

See Watanabe [W] for the precise conditions. So λ takes the place of $\frac{d}{2}$, and it controls both the free energy and the generalization error. The expansion localizes: restricting the integral to a neighborhood of a point of W_0 gives the same expansion with that neighborhood's own learning coefficient. Hence, as n grows, the posterior concentrates on the regions of W_0 with the smallest learning coefficient. This is the precise sense of the slogan that learning prefers “simpler” solutions: simplicity means a smaller learning coefficient, not fewer parameters.

Example (Selection by volume). Let $W = [-1, 2]$ with the uniform prior, and let

$$L(w) = w^2(w - 1)^4.$$

Then $W_0 = \{0, 1\}$, and L behaves like w^2 near 0 and like $(w - 1)^4$ near 1. So $\lambda = \frac{1}{2}$ near 0 and $\lambda = \frac{1}{4}$ near 1. With L in place of L_n , Laplace's method gives

$$\int_{-1}^{\frac{1}{2}} e^{-nL(w)} dw = \sqrt{\frac{\pi}{n}} (1 + o(1)) \quad \text{and} \quad \int_{\frac{1}{2}}^2 e^{-nL(w)} dw = 2\Gamma\left(\frac{5}{4}\right) n^{-\frac{1}{4}} (1 + o(1)).$$

So the posterior odds of the quartic well against the quadratic well are

$$\frac{\Gamma\left(\frac{1}{4}\right)}{2\sqrt{\pi}} n^{\frac{1}{4}} (1 + o(1)) \rightarrow \infty.$$

The reason is volume: near 1, the set $\{L < \varepsilon\}$ has length about $2\varepsilon^{\frac{1}{4}}$, much larger than the length about $2\varepsilon^{\frac{1}{2}}$ near 0. In general, let U_1 and U_2 be neighborhoods of two parts of W_0 on which the volume functions satisfy volume laws with the exponents (λ_1, m_1) and (λ_2, m_2) and the constants c_1 and c_2 . Then Proposition 6.2 gives

$$\frac{\int_{U_1} \varphi e^{-nL} dw}{\int_{U_2} \varphi e^{-nL} dw} = \frac{\Gamma(\lambda_1 + 1) c_1}{\Gamma(\lambda_2 + 1) c_2} n^{\lambda_2 - \lambda_1} (\log n)^{m_1 - m_2} (1 + o(1)).$$

So the region with the smaller learning coefficient carries more posterior mass by a power of n , and the constants matter only when the exponents agree. If $\lambda_1 < \lambda_2$ and $L - \delta$ satisfies the volume law on U_1 for some $\delta > 0$, then the ratio acquires the factor $e^{-n\delta}$. This is the trade-off between loss and learning coefficient in the remark on phase transitions in Chapter 5. With $\frac{1}{\varepsilon}$ in place of n , the same computation compares the Gibbs masses of two wells (Proposition 9.1).

Remark (The random term). A sequence of random variables T_n is bounded in probability, written $T_n = O_p(1)$, if for every $\eta > 0$ there is $C > 0$ such that $\mathbb{P}(|T_n| > C) < \eta$ for all n . It tends to 0 in probability, written $T_n = o_p(1)$, if $\mathbb{P}(|T_n| > \eta) \rightarrow 0$ for every $\eta > 0$. We explain why the random part of F_n is $O_p(1)$ but not $o_p(1)$.

Consider first the regular case of Proposition 6.1. Write

$$s(x) = \nabla_w \log p(x | w_0) \quad \text{and} \quad J(x) = -\nabla_w^2 \log p(x | w_0)$$

for the score and minus its derivative. Then $\mathbb{E}[s] = 0$ and $\mathbb{E}[ss^t] = \mathbb{E}[J] = I$, by Section 1.4 and the proof of Proposition 1.3. With $v = \sqrt{n}(w - w_0)$, Taylor's formula gives

$$nL_n(w) = \frac{1}{2} v^t I_n v - v^t Y_n + O\left(\frac{|v|^3}{\sqrt{n}}\right),$$

where

$$Y_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n s(X_j) \quad \text{and} \quad I_n = \frac{1}{n} \sum_{j=1}^n J(X_j).$$

By the central limit theorem, Y_n converges in distribution to a Gaussian vector with covariance I . By the law of large numbers, $I_n \rightarrow I$. Completing the square,

$$\frac{1}{2} v^t I v - v^t Y_n = \frac{1}{2} (v - I^{-1} Y_n)^t I (v - I^{-1} Y_n) - \frac{1}{2} Y_n^t I^{-1} Y_n.$$

So the data shift the Gaussian by $n^{-\frac{1}{2}} I^{-1} Y_n$, which does not change its integral, and they lower its minimum by $\frac{1}{2} Y_n^t I^{-1} Y_n$. Hence

$$F_n = n S_n - \log Z(n) - \frac{1}{2} Y_n^t I^{-1} Y_n + o_p(1).$$

The random term converges in distribution to $-\frac{1}{2}|G|^2$, where G is a standard Gaussian vector in $\mathbb{R}^d$. The shifted Gaussian is the Bernstein-von Mises approximation of Section 1.5.

In a singular model, the same mechanism works in the coordinates u of a resolution of singularities. Write $u^k = u_1^{k_1} \cdots u_d^{k_d}$. In a chart in which $L(g(u)) = u^{2k}$, Watanabe [W] shows that

$$n L_n(g(u)) = n u^{2k} - \sqrt{n} u^k \xi_n(u),$$

where ξ_n is a random function which converges in distribution to a Gaussian process. With $t = n u^{2k}$ and $u_j > 0$, this reads

$$n L_n(g(u)) = t - \sqrt{t} \xi_n(u).$$

So on the region $t \lesssim 1$, which carries the integral, the random term is as large as the deterministic one. Hence $\int_W \varphi e^{-n L_n} dw$ is $Z(n)$ times a random factor which is bounded and bounded away from 0 in probability, but which does not tend to 1.

Remark (Free energy and generalization error). The two statements of Watanabe's theorem are linked by an exact identity. Recall that the Bayesian predictive distribution is

$$p_n(x) = \int_W p(x | w) \pi_n(dw),$$

as in Section 1.5. Since $p(x | w) \prod_{j=1}^n p(X_j | w)$ is the likelihood of the $n+1$ samples $X_1, \dots, X_n, x$, we have

$$p_n(X_{n+1}) = \frac{Z_{n+1}}{Z_n}, \quad \text{that is,} \quad -\log p_n(X_{n+1}) = F_{n+1} - F_n.$$

Let $S = -\int q_0 \log q_0$ be the entropy of q_0 . Taking expectations, we obtain

$$\mathbb{E}[\text{KL}(q_0 \| p_n)] = \mathbb{E}[F_{n+1}] - \mathbb{E}[F_n] - S.$$

So the expected generalization error is the increment of the expected free energy, minus the entropy. Since $F_0 = -\log \int_W \varphi = 0$, summation gives

$$\sum_{j=0}^{n-1} \mathbb{E}[\text{KL}(q_0 \| p_j)] = \mathbb{E}[F_n] - nS = \lambda \log n + o(\log n),$$

by the expansion of the expected free energy in Watanabe's theory [W]. So, if $n\mathbb{E}[\text{KL}(q_0 \| p_n)]$ has a limit as $n \rightarrow \infty$, then this limit is λ .

The regular case is the least favorable one.

Lemma 6.4. *Suppose that $L \in C^2(W)$ and W_0 contains an interior point of W . Then $\lambda \leq \frac{d}{2}$.*

PROOF. Let $w_* \in W_0$ be an interior point. Since L attains its minimum 0 at w_* , we have $\nabla L(w_*) = 0$. So Taylor's formula gives

$$L(w) \leq \Lambda |w - w_*|^2 \quad \text{with } \Lambda = \frac{1}{2} \max \|\nabla^2 L\|$$

near w_* . Hence,

$$V(\varepsilon) \geq (\min \varphi) \omega_d \left(\frac{\varepsilon}{\Lambda} \right)^{\frac{d}{2}}$$

for small ε . If $\lambda > \frac{d}{2}$, then the volume law gives

$$V(\varepsilon) \varepsilon^{-\frac{d}{2}} \rightarrow 0,$$

a contradiction. $\square$

Remark (Flatness and the Hessian). In machine learning, a minimum is often called flat if the Hessian of the loss has small eigenvalues. The learning coefficient measures flatness at all orders. Suppose that L is real-analytic. Let $w_* \in W_0$ be an interior point, let $\lambda(w_*)$ be the learning coefficient of a small ball around w_* (the local learning coefficient of the next section), and let r be the rank of $\nabla^2 L(w_*)$. Then

$$\frac{r}{2} \leq \lambda(w_*) \leq \frac{d}{2}.$$

The upper bound is Lemma 6.4 applied to the ball. For the lower bound, write $w - w_* = (u, v)$ with u in the range and v in the kernel of $\nabla^2 L(w_*)$. Since $\partial_u^2 L(w_*)$ is positive definite, the implicit function theorem gives a smooth map ψ such that

$$\psi(0) = 0 \quad \text{and} \quad \partial_u L(\psi(v), v) = 0.$$

By Taylor's formula in u , there is $c > 0$ such that

$$L(u, v) \geq L(\psi(v), v) + c|u - \psi(v)|^2 \geq c|u - \psi(v)|^2$$

near w_* . So near w_* , the set $\{L < \varepsilon\}$ lies in the tube $\{|u - \psi(v)| < (\frac{\varepsilon}{c})^{\frac{1}{2}}\}$, whose volume is $O(\varepsilon^{\frac{r}{2}})$. Hence $\lambda(w_*) \geq \frac{r}{2}$. Both bounds can be strict: for $L(x, y) = x^2 + y^4$, the scaling $(x, y) \mapsto (\varepsilon^{\frac{1}{2}}x, \varepsilon^{\frac{1}{4}}y)$ gives $\lambda(0) = \frac{3}{4}$, while $r = 1$ and $d = 2$. On the other hand, the size of the non-zero eigenvalues does not affect λ : for $L(x, y) = x^2 + \kappa y^2$, we have $\lambda(0) = 1$ for every $\kappa > 0$, and κ enters only the constant of the volume law. So a flatter region in the sense of the learning coefficient has more degenerate directions, counted with their orders of vanishing, and not merely a smaller curvature. Both notions appear in Chapter 10. There, the strata of W_0 differ in λ , and along a stratum the Langevin dynamics drifts toward smaller transversal curvature, which changes only the constant of the volume law.

6.4. The local learning coefficient

The learning coefficient of a region is the minimum of local invariants attached to its points.

Definition (Local learning coefficient). Let L be real-analytic near w_* and not locally constant. The local learning coefficient $\lambda(w_*)$ and the multiplicity $m(w_*)$ are defined by the volume law

$$\text{Vol}(\{w \in B(w_*, \delta) : |L(w) - L(w_*)| < \varepsilon\}) = c \varepsilon^{\lambda(w_*)} \left(\log \frac{1}{\varepsilon} \right)^{m(w_*)-1} (1 + o(1))$$

for small $\delta > 0$. At a local minimum, the absolute value is not needed, and the definition is that of Lau et al. [LFW]. At a saddle point, this two-sided definition is one of several possible choices.

Example.

- (1) For $L(x, y) = x^2 + y^2$, $\lambda(0) = 1$ and $m(0) = 1$.
- (2) For $L(x, y) = x^2 y^2$, $(\lambda, m) = (\frac{1}{2}, 2)$ at the origin and $(\frac{1}{2}, 1)$ at the other points of the axes.
- (3) For the saddle $L(x, y) = x^2 - y^2$, the two-sided definition gives $(\lambda, m) = (1, 2)$ at the origin.

In (3), the substitution $u = x - y$ and $v = x + y$ turns $\{|x^2 - y^2| < \varepsilon\}$ into $\{|uv| < \varepsilon\}$, a region of the same shape as in (2), which explains the logarithm. The oscillatory integral of Section 6.3 sees this saddle differently. By stationary phase, for φ smooth with compact support,

$$\int_{\mathbb{R}^2} e^{i\tau(x^2 - y^2)} \varphi(x, y) dx dy = \frac{\pi}{\tau} \varphi(0, 0) + O(\tau^{-2}) \quad \text{as } \tau \rightarrow \infty,$$

with the decay rate τ^{-1} of a volume law with $\lambda = 1$, but without the factor $\log \tau$. For this saddle, the logarithmic contributions of $\{L > 0\}$ and $\{L < 0\}$ cancel, and the oscillatory integral and the two-sided volume no longer carry the same information. This is one reason why there is no canonical invariant at a saddle point; we return to it in Chapter 7.

Lemma 6.5. *Let L be C^1 near w_* , and suppose that $L - L(w_*)$ takes both signs in every neighborhood of w_* . Then for every sufficiently small $\delta > 0$, there is $c > 0$ such that*

$$\text{Vol}(\{w \in B(w_*, \delta) : |L(w) - L(w_*)| < \varepsilon\}) \geq c\varepsilon \quad \text{for } 0 < \varepsilon \leq 1.$$

In particular, if L is real-analytic, then the two-sided local learning coefficient satisfies $\lambda(w_) \leq 1$.*

PROOF. Write $f = L - L(w_*)$. Choose points $p_{\pm} \in B(w_*, \frac{\delta}{2})$ and $\rho > 0$ such that $B(p_{\pm}, \rho) \subset B(w_*, \frac{\delta}{2})$ and $\pm f > 0$ on $B(p_{\pm}, \rho)$. Let $e = p_- - p_+$, and let D be the disc of radius ρ centered at p_+ in the hyperplane through p_+ orthogonal to e . For $q \in D$, the function $t \mapsto f(q + te)$ is positive at $t = 0$ and negative at $t = 1$, so it vanishes at some $t(q) \in (0, 1)$. Hence

$$|f(q + te)| \leq \Lambda |e| |t - t(q)| \quad \text{with } \Lambda = \sup_{B(w_*, \delta)} |\nabla f|.$$

So the segment $\{q + te : 0 \leq t \leq 1\}$, which lies in $B(w_*, \frac{\delta}{2})$, meets $\{|f| < \varepsilon\}$ in a set of length at least $\min\{\frac{\varepsilon}{\Lambda}, |e|\}$. These segments are disjoint, and Fubini's theorem gives the lower bound with

$$c = \text{Vol}_{d-1}(D) \min \left\{ \frac{1}{\Lambda}, |e| \right\}.$$

□

The logarithm at the saddle $x^2 - y^2$ is special to dimension two. Let $Q(x, y) = |x|^2 - |y|^2$ with $x \in \mathbb{R}^p$, $y \in \mathbb{R}^q$, and $p, q \geq 1$. Polar coordinates in x and in y , followed by the substitution $a = |x|^2$ and $b = |y|^2$, give

$$\text{Vol}(\{(x, y) : |x| < \delta, |y| < \delta, |Q(x, y)| < \varepsilon\}) = \frac{|\mathbb{S}^{p-1}| |\mathbb{S}^{q-1}|}{4} \int_0^{\delta^2} \int_0^{\delta^2} \mathbb{1}_{\{|a-b| < \varepsilon\}} a^{\frac{p}{2}-1} b^{\frac{q}{2}-1} da db.$$

For $p = q = 1$, the right side is

$$2\varepsilon \log \frac{1}{\varepsilon} (1 + o(1)).$$

For $p + q \geq 3$, the function $a^{\frac{p+q}{2}-2}$ is integrable near 0, and the right side is comparable to ε . So, by the Morse lemma, a non-degenerate saddle point in $d \geq 3$ variables has the two-sided invariants $(1, 1)$.

At points of W_0 , the local invariants can often be computed without a resolution of singularities, by the following rules.

Lemma 6.6. *Let L , L_1 , and L_2 be non-negative real-analytic functions near w_* , x_* , and y_* , respectively, which vanish at these points and are not identically zero. Suppose that L_1 and L_2 have the local invariants (λ_1, m_1) at x_* and (λ_2, m_2) at y_* , and that $\lambda_1 \leq \lambda_2$.*

- (1) *Comparison: If $L' \geq 0$ is real-analytic and $c_1 L \leq L' \leq c_2 L$ near w_* for some constants $0 < c_1 \leq c_2$, then L' has the same local invariants at w_* as L . If Ψ is a real-analytic diffeomorphism, then $L \circ \Psi$ has the same local invariants at $\Psi^{-1}(w_*)$ as L at w_* .*
- (2) *Free variables: The function $(x, y) \mapsto L_1(x)$ has the local invariants (λ_1, m_1) at (x_*, y_*) .*
- (3) *Sums: The function $L_1(x) + L_2(y)$ has the local invariants $(\lambda_1 + \lambda_2, m_1 + m_2 - 1)$ at (x_*, y_*) .*
- (4) *Products: The function $L_1(x)L_2(y)$ has the local invariants (λ_1, m_1) at (x_*, y_*) if $\lambda_1 < \lambda_2$, and $(\lambda_1, m_1 + m_2)$ if $\lambda_1 = \lambda_2$.*

PROOF. We may replace the balls in the definition by products of balls. In (1),

$$\left\{ L < \frac{\varepsilon}{c_2} \right\} \subset \{ L' < \varepsilon \} \subset \left\{ L < \frac{\varepsilon}{c_1} \right\}.$$

By the volume law of L , the volumes of the two outer sets are comparable to $\varepsilon^\lambda (\log \frac{1}{\varepsilon})^{m-1}$. Since L' satisfies a volume law as well, its exponents are those of L . A diffeomorphism changes volumes by a bounded factor and maps small balls into balls. Part (2) is Fubini's theorem. In (3),

$$\left\{ L_1 < \frac{\varepsilon}{2} \right\} \times \left\{ L_2 < \frac{\varepsilon}{2} \right\} \subset \{ L_1 + L_2 < \varepsilon \} \subset \{ L_1 < \varepsilon \} \times \{ L_2 < \varepsilon \}.$$

So the volume function of $L_1 + L_2$ is comparable to

$$\varepsilon^{\lambda_1 + \lambda_2} \left(\log \frac{1}{\varepsilon} \right)^{m_1 + m_2 - 2},$$

and since it satisfies a volume law, its exponents are $\lambda_1 + \lambda_2$ and $m_1 + m_2 - 1$. In (4), Fubini's theorem gives $\zeta = \zeta_1 \zeta_2$ for the local zeta functions. The largest pole of ζ_i is $-\lambda_i$, of order m_i , and $\zeta_i(z) > 0$ for real $z > -\lambda_i$. So (4) follows from the remark on the zeta function in Section 6.3. $\square$

For instance, the invariants of $x^2 y^2$ in the example above follow from (4) at the origin, and from (1) and (2) at the other points of the axes. For $a \in \mathbb{R}^q$, the function $|a|^2$ has the invariants $(\frac{q}{2}, 1)$ at 0, by (3). So for $a \in \mathbb{R}^q$ and $b \in \mathbb{R}^p$, the function $|a|^2 |b|^2$ has the invariants $(\frac{\min\{p, q\}}{2}, 1)$ at the origin if $p \neq q$, and $(\frac{p}{2}, 2)$ if $p = q$. For the two-layer linear network with $H = 1$ and $\Phi = 0$, we have $A = a^t$ and $B = b$ with $a \in \mathbb{R}^M$ and $b \in \mathbb{R}^N$, and

$$L = \frac{1}{2} |a|^2 |b|^2.$$

So L has the invariants $(\frac{\min\{M, N\}}{2}, 1)$ at the origin if $M \neq N$, and $(\frac{M}{2}, 2)$ if $M = N$. These are the values of the Aoyagi-Watanabe theorem of Section 6.6 for $H = 1$ and $r = 0$.

The local learning coefficient is estimated in practice by sampling. With a tempering parameter $\beta > 0$, the tempered posterior is proportional to $\varphi(w) e^{-\beta n L_n(w)}$. At $\beta = \frac{1}{\log n}$, the average of $n L_n$ over the tempered posterior exceeds its minimum by $\lambda \log n$ to leading order, by a theorem of Watanabeⁱ. So λ can be estimated by sampling, with the Langevin sampler of Section 3.3, without knowing q_0 or resolving any singularity. With L in place of L_n , the theorem rests on the following identity, in which N plays the role of βn .

ⁱS. Watanabe, *A widely applicable Bayesian information criterion*. J. Mach. Learn. Res. 14 (2013), 867–897.

Proposition 6.7. *Suppose that V satisfies the volume law of Proposition 6.2. For $N > 0$, let $\mathbb{E}_N$ be the expectation with respect to the Gibbs measure*

$$\frac{1}{Z(N)} \varphi(w) e^{-NL(w)} dw$$

on W . Then

$$N\mathbb{E}_N[L] = -\frac{d \log Z(N)}{d \log N} \rightarrow \lambda \quad \text{as } N \rightarrow \infty.$$

PROOF. The identity follows by differentiation under the integral sign. Since

$$\int_s^\infty (u-1)e^{-u} du = se^{-s} \quad \text{for } s \geq 0,$$

we have

$$NL(w)e^{-NL(w)} = \int_0^\infty (u-1)e^{-u} \mathbf{1}_{\{NL(w) < u\}} du.$$

So Fubini's theorem gives

$$\int_W NL(w) \varphi(w) e^{-NL(w)} dw = \int_0^\infty (u-1)e^{-u} V\left(\frac{u}{N}\right) du.$$

The proof of Proposition 6.2, with the weight $(u-1)e^{-u}$ in place of e^{-u} and with the dominating function $|u-1|e^{-u}$, shows that this integral is

$$(\Gamma(\lambda+2) - \Gamma(\lambda+1)) cN^{-\lambda} (\log N)^{m-1} (1 + o(1)) = \lambda \Gamma(\lambda+1) cN^{-\lambda} (\log N)^{m-1} (1 + o(1)).$$

By Proposition 6.2,

$$Z(N) = \Gamma(\lambda+1) cN^{-\lambda} (\log N)^{m-1} (1 + o(1)),$$

and dividing completes the proof. $\square$

In the regular case, $\lambda = \frac{d}{2}$, so each of the d quadratic directions carries the mean energy $\frac{1}{2N}$, as in the equipartition of energy in statistical mechanics. The convergence is slow. For $L(a, b) = a^2 b^2$ on $[-1, 1]^2$ with $\varphi = \frac{1}{4}$, the volume function computed in Section 6.3 gives

$$N\mathbb{E}_N[L] = \frac{1}{2} - \frac{1}{\log N} + O\left(\frac{1}{(\log N)^2}\right),$$

which is the derivative of $\lambda \log N - (m-1) \log \log N$ with respect to $\log N$, up to the error term.

To estimate the local coefficient at a point w_* , Lau et al. [LFW] pin the posterior near w_* by a Gaussian factor $e^{-\frac{\gamma}{2}|w-w_*|^2}$. Let $\mathbb{E}_{n,\gamma}$ be the expectation with respect to the probability measure on W with density proportional to

$$\exp\left(-\frac{nL_n(w)}{\log n} - \frac{\gamma}{2}|w-w_*|^2\right).$$

Their estimator is

$$\hat{\lambda}_n(w_*; \gamma) = \frac{\mathbb{E}_{n,\gamma}[nL_n] - nL_n(w_*)}{\log n}.$$

In Problem 6.1, the expectation is taken to be exact. If L_n is replaced by L and $w_* \in W_0$, then $\hat{\lambda}_n$ is $N\mathbb{E}_N[L]$ with $N = \frac{n}{\log n}$, for the prior proportional to $e^{-\frac{\gamma}{2}|w-w_*|^2}$ on W . If, moreover, W is the disjoint union of U_1 and U_2 , then

$$N\mathbb{E}_N[L] = \frac{Z_1(N)}{Z(N)} N\mathbb{E}_{N,1}[L] + \frac{Z_2(N)}{Z(N)} N\mathbb{E}_{N,2}[L],$$

where Z_i and $\mathbb{E}_{N,i}$ are defined with the integrals restricted to U_i . So, by Proposition 6.7 applied to each U_i , the estimator is approximately the average of the learning coefficients of U_1 and U_2 , weighted by their posterior masses. The following problem asks when it localizes.

Problem 6.1 (MAIS-O69). *Let $W = \{w \in \mathbb{R}^2 : |w| \leq 3\}$. Consider the model $y \sim \mathcal{N}(\mu(w), 1)$ with*

$$\mu(w) = (w_1 - 1) (w_1^2 + w_2^2)^2,$$

and the truth $y \sim \mathcal{N}(0, 1)$. Then $L(w) = \frac{1}{2}\mu(w)^2$ and $W_0 = \{w_1 = 1\} \cup \{(0, 0)\}$, with $\lambda = \frac{1}{2}$ on the segment and $\lambda = \frac{1}{4}$ at the origin. Let $w_ = (1, 0)$, and let $\hat{\lambda}_n(w_*; \gamma)$ be the estimator of [LFW] with the tempering $\beta = \frac{1}{\log n}$ and the localization strength γ . Determine the sequences $\gamma_n \rightarrow \infty$ for which $\hat{\lambda}_n(w_*; \gamma_n) \rightarrow \frac{1}{2}$ in probability, and identify the limit at the transitional scaling.*

For fixed γ , the Gaussian factor is just another prior on W , and the estimator converges to the global value $\frac{1}{4}$, as in Proposition 6.7. Under a hard restriction to a small ball around w_* , it converges to $\frac{1}{2}$. Problem MAIS-O69 asks for the boundary between the two regimes. By the decomposition above, it is a two-well problem for Laplace's method, with the large parameter $N = \frac{n}{\log n}$. Heuristically, the posterior mass near the origin is of order $e^{-\frac{\gamma}{2}} N^{-\frac{1}{4}}$, and the mass near w_* is of order $\gamma^{-\frac{1}{2}} N^{-\frac{1}{2}}$. This suggests that the transition occurs at $\gamma_n \sim \frac{1}{2} \log n$.

6.5. The stratification by the learning coefficient

On the optimal set W_0 , the local learning coefficient is a function $\lambda : W_0 \rightarrow \mathbb{Q}_{>0}$. Any theory of effective dynamics on the strata of W_0 needs some regularity of the level sets of λ , such as the one asked for in Question 6.1 below.

Example. On $\mathbb{T}^2$, let $L(x, y) = \sin^2 x \sin^4 y$. Then W_0 is the union of the two vertical circles $\{x \in \{0, \pi\}\}$ and the two horizontal circles $\{y \in \{0, \pi\}\}$. Off the crossings, $\lambda = \frac{1}{2}$ on the vertical circles (quadratic transversal profile) and $\lambda = \frac{1}{4}$ on the horizontal circles (quartic transversal profile). At the four crossings, $\lambda = \frac{1}{4}$ and $m = 1$, by Lemma 6.6(4). So the stratification has two pieces, and the crossings belong to the flatter one. This is the example of MAIS-O75 (Chapter 10).

Question 6.1 (MAIS-O72). *Let $L : \mathbb{T}^d \rightarrow [0, \infty)$ be real-analytic and not identically zero, with $W_0 = L^{-1}(0) \neq \emptyset$. For $c > 0$, let*

$$S_c = \{w \in W_0 : \lambda(w) = c\}.$$

- (a) *Does λ take only finitely many values on W_0 ?*
- (b) *Is each S_c a locally closed subanalytic subset of $\mathbb{T}^d$, and does the frontier condition $\overline{S_c} \setminus S_c \subset \bigcup_{c' < c} S_{c'}$ hold?*

The same questions are asked for the pair (λ, m) , ordered by declaring (λ', m') deeper than (λ, m) if $\lambda' < \lambda$, or $\lambda' = \lambda$ and $m' > m$.

Remark (Subanalytic sets and resolutions). We recall the notions in Question 6.1. A subset of $\mathbb{T}^d$ is semianalytic if each point of $\mathbb{T}^d$ has a neighborhood U in which the set is a finite union of sets of the form

$$\{w \in U : f(w) = 0, g_1(w) > 0, \dots, g_l(w) > 0\}$$

with real-analytic functions f and g_i on U . It is subanalytic if each point has a neighborhood U in which the set is the projection to U of a bounded semianalytic subset of $U \times \mathbb{R}^e$ for some e . The image of a semianalytic set under a proper real-analytic map is subanalytic, but in general

not semianalytic. Finite unions, finite intersections, complements, and closures of subanalytic sets are subanalytic; see Bierstone-Milmanⁱ. A set is locally closed if it is the intersection of an open set and a closed set, i.e., if it is open in its closure. The frontier condition says that S_c is bordered only by points of smaller learning coefficient. In the example on $\mathbb{T}^2$ above, the four crossings border the vertical circles, and they have the smaller coefficient $\frac{1}{4}$.

It is known that λ is lower semicontinuous on W_0 , i.e., that $\{\lambda \leq c\}$ is closed for every c , by a result of Lehalleur-Rimányiⁱⁱ. The following sketch indicates that (a) and (b) follow from the resolution of singularities. Let g be a resolution as in the remark on the zeta function, over a neighborhood of a point of W_0 . For a point p with $L(g(p)) = 0$, choose a chart around p in which

$$L(g(u)) = u_1^{2k_1} \cdots u_d^{2k_d} \quad \text{and} \quad |\det g'(u)| = b(u) |u_1^{h_1} \cdots u_d^{h_d}|,$$

and set

$$\nu(p) = \min \left\{ \frac{h_j + 1}{2k_j} : u_j(p) = 0, k_j > 0 \right\}.$$

In a small box around p , the factors $u_j^{2k_j}$ and $|u_j|^{h_j}$ with $u_j(p) \neq 0$ are bounded above and below. So, as in the remark on the zeta function, the largest pole of the contribution of this box to the zeta function is $-\nu(p)$. In particular, $\nu(p)$ does not depend on the chart. For $w \in W_0$ near the point, the fiber $g^{-1}(w)$ is compact, since g is proper. For small δ , the set $g^{-1}(B(w, \delta))$ lies in a small neighborhood of $g^{-1}(w)$, which we cover by finitely many small boxes centered at points of $g^{-1}(w)$, with a partition of unity. The contributions of these boxes to the local zeta function at w are positive for real z , so their poles do not cancel. Hence

$$\lambda(w) = \min \left\{ \nu(p) : p \in g^{-1}(w) \right\}.$$

A point near p lies only on hyperplanes $\{u_j = 0\}$ which pass through p . So ν is lower semicontinuous, and $\{\nu \leq c\}$ is locally a finite union of such hyperplanes. Since g is proper, $\{\lambda \leq c\} = g(\{\nu \leq c\})$ is closed and subanalytic near the point. This recovers the lower semicontinuity of λ .

Moreover, W_0 is compact, so it is covered by finitely many such neighborhoods, and, since g is proper, finitely many charts suffice over each of them. So ν takes only finitely many values, and hence so does λ , which is (a). Then $\{\lambda < c\} = \{\lambda \leq c'\}$ for the largest value $c' < c$ of λ , so $\{\lambda < c\}$ is closed. Hence $S_c = \{\lambda \leq c\} \setminus \{\lambda < c\}$ is locally closed and subanalytic, and

$$\overline{S_c} \setminus S_c \subset \{\lambda \leq c\} \setminus S_c = \{\lambda < c\},$$

which is (b). For the pair (λ, m) , the multiplicity $m(w)$ is the largest number of indices j with $u_j(p) = 0$ and $\frac{h_j+1}{2k_j} = \lambda(w)$, over the points $p \in g^{-1}(w)$. The same idea applies, with intersections of the hyperplanes $\{u_j = 0\}$ in place of single hyperplanes, but the orders of the poles have to be followed through the sum over the boxes, and this refined version needs more care.

6.6. Computing learning coefficients

Learning coefficients are computed by resolving singularities, which is feasible for small models. The first nontrivial model computed in this way is reduced rank regression, i.e., the two-layer linear network of Section 3.5 viewed as the statistical model $y = BAx + \xi$ with Gaussian noise.

ⁱE. Bierstone and P. Milman, *Semianalytic and subanalytic sets*. Inst. Hautes Études Sci. Publ. Math. 67 (1988), 5–42.

ⁱⁱS. Lehalleur and R. Rimányi, *Geometry of fibers of the multiplication map of deep linear neural networks*. arXiv:2411.19920.

Aoyagi-Watanabeⁱ computed its learning coefficient by a resolution of singularities carried out by hand.

Theorem (Aoyagi-Watanabe). *Consider reduced rank regression with input dimension M , output dimension N , width H , and true rank r .*

(1) *If $N + r \leq M + H$, $M + r \leq N + H$, and $H + r \leq M + N$, then*

$$\lambda = \frac{2(H + r)(M + N) - (M - N)^2 - (H + r)^2 + \delta}{8},$$

where $\delta = 1$ if $M + H + N + r$ is odd and $\delta = 0$ otherwise. Moreover, $m = 2$ if $M + H + N + r$ is odd and $m = 1$ otherwise.

(2) *If $M + H < N + r$, then $\lambda = \frac{1}{2}(HM - Hr + Nr)$ and $m = 1$.*

(3) *If $N + H < M + r$, then $\lambda = \frac{1}{2}(HN - Hr + Mr)$ and $m = 1$.*

(4) *If $M + N < H + r$, then $\lambda = \frac{1}{2}MN$ and $m = 1$.*

In the case $M = N = H$, only case (1) occurs, and we write λ_r^{AW} for the learning coefficient with true rank r . Then

$$\lambda_r^{\text{AW}} = \frac{(H + r)(3H - r) + \delta_r}{8} \quad \text{and} \quad \lambda_r^{\text{AW}} - \lambda_{r-1}^{\text{AW}} = \frac{2(H - r) + 1 + \delta_r - \delta_{r-1}}{8},$$

where $\delta_r = 1$ if $3H + r$ is odd and $\delta_r = 0$ otherwise. So the learning coefficient strictly increases with the rank r of the target as long as $r < H$. This global value is attained at the most degenerate points of the fiber F_Φ . The full table of local invariants is open.

Remark (A conservation law). In the deep linear network (3.1), the symmetry of Proposition 2.1 gives a conservation law, which constrains the point of F_Φ that gradient flow reaches. Write $R = BA - \Phi$, so that $\dot{A} = -B^t R$ and $\dot{B} = -RA^t$. Then

$$\frac{d}{dt} (AA^t) = -B^t RA^t - AR^t B = \frac{d}{dt} (B^t B).$$

So $AA^t - B^t B$ is constant along gradient flow. In the proof of Theorem 3.3, this is the conservation of $c_\alpha^2 - d_\alpha^2$. Suppose that the initialization is balanced, i.e., $AA^t = B^t B$ at $t = 0$, as in Theorem 3.3. If the flow converges to a point $(A, B) \in F_\Phi$, then $AA^t = B^t B$ there, so

$$\Phi^t \Phi = A^t (B^t B) A = (A^t A)^2.$$

Since $A^t A$ is symmetric, it has the same rank as its square. Hence

$$\text{rank } A = \text{rank } (A^t A) = \text{rank } (\Phi^t \Phi) = r \quad \text{and} \quad \text{rank } B = \text{rank } (B^t B) = \text{rank } (AA^t) = r.$$

So the flow selects the most degenerate points of F_Φ , those with $\text{rank } A = \text{rank } B = r$, at which the local learning coefficient is the Aoyagi-Watanabe coefficient. Hence, in this model, if gradient flow from a balanced initialization converges to a point of F_Φ , then this point has the minimal local learning coefficient, which the Bayesian selection prefers.

At less degenerate points of F_Φ , the local learning coefficient can be larger. For instance, suppose that $\text{rank } A = M$ or $\text{rank } B = N$ at $w = (A, B) \in F_\Phi$. Then the differential $(\dot{A}, \dot{B}) \mapsto \dot{B}A + B\dot{A}$ of the multiplication map $(A, B) \mapsto BA$ is onto $M_{N,M}(\mathbb{R})$. So near w , the fiber F_Φ is a submanifold of codimension MN , and L is half the squared norm of MN local coordinates. Hence $\lambda(w) = \frac{MN}{2}$ and $m(w) = 1$, by Lemma 6.6. For $M = N = H = 2$ and $\Phi = 0$, this gives $\lambda(w) = 2$ at $w = (I, 0)$, while the Aoyagi-Watanabe theorem gives $\lambda = \frac{3}{2}$.

ⁱM. Aoyagi and S. Watanabe, *Stochastic complexities of reduced rank regression in Bayesian estimation*. Neural Networks 18 (2005), 924–933.

The local invariants are constant on large subsets of F_Φ , by linear algebra. The group $\text{GL}(H, \mathbb{R})$ acts by $(A, B) \mapsto (gA, Bg^{-1})$ and preserves L . The pairs $(h, k) \in \text{GL}(M, \mathbb{R}) \times \text{GL}(N, \mathbb{R})$ with $k\Phi h^{-1} = \Phi$ act by $(A, B) \mapsto (Ah^{-1}, kB)$. They preserve F_Φ and change L into $\frac{1}{2}\|k(BA - \Phi)h^{-1}\|_F^2$, which lies between two positive multiples of L . By Lemma 6.6(1), such linear changes of variables do not change $(\lambda(w), m(w))$, and neither does a fixed change of bases in $\mathbb{R}^M$ and $\mathbb{R}^N$, which replaces Φ by an equivalent matrix. So we may assume that

$$\Phi = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Suppose that $\text{rank } A = a$ and $\text{rank } B = b$. Then $BA = \Phi$ forces $\dim(\text{im } A \cap \ker B) = a - r$. Choosing bases adapted to $\ker A \subset \ker \Phi \subset \mathbb{R}^M$, to $\text{im } A \cap \ker B$, $\text{im } A$, and $\ker B$ in $\mathbb{R}^H$, and to $\text{im } \Phi \subset \text{im } B \subset \mathbb{R}^N$ shows that (A, B) lies in the orbit of the normal form

$$A_{a,b} = \begin{pmatrix} I_r & 0 & 0 \\ 0 & I_{a-r} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad B_{a,b} = \begin{pmatrix} I_r & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{b-r} \\ 0 & 0 & 0 & 0 \end{pmatrix},$$

with the blocks of sizes $(r, a - r, M - a)$ in $\mathbb{R}^M$, $(r, a - r, H - a - b + r, b - r)$ in $\mathbb{R}^H$, and $(r, b - r, N - b)$ in $\mathbb{R}^N$. So the orbits of the combined action on F_Φ are exactly the sets

$$\{(A, B) \in F_\Phi : \text{rank } A = a, \text{rank } B = b\},$$

where $r \leq a \leq \min\{H, M\}$, $r \leq b \leq \min\{H, N\}$, and $a + b \leq H + r$. The last inequality is Sylvester's rank inequality $\text{rank}(BA) \geq \text{rank } A + \text{rank } B - H$. Hence $(\lambda(w), m(w))$ depends only on the pair $(\text{rank } A, \text{rank } B)$. The following problem asks for the table. Some entries, such as those with $\text{rank } A = M$ or $\text{rank } B = N$ above, follow from these normal forms and the rules of Lemma 6.6. At $(a, b) = (r, r)$, $\lambda(w)$ is the Aoyagi-Watanabe coefficient, whose computation required a resolution of singularities. The table is the first step toward MAIS-O77 below.

Problem 6.2 (MAIS-O70). *Consider the loss $L(A, B) = \frac{1}{2}\|BA - \Phi\|_F^2$ of reduced rank regression, where $A \in M_{H,M}(\mathbb{R})$, $B \in M_{N,H}(\mathbb{R})$, and $\Phi \in M_{N,M}(\mathbb{R})$ has rank $r \leq H$. For every pair of integers (a, b) with $r \leq a \leq \min\{H, M\}$, $r \leq b \leq \min\{H, N\}$, and $a + b \leq H + r$, compute the local invariants $(\lambda(w), m(w))$ at the points $w = (A, B) \in F_\Phi$ with $\text{rank } A = a$ and $\text{rank } B = b$. Determine the pairs (a, b) at which $\lambda(w)$ equals the Aoyagi-Watanabe coefficient λ .*

A natural first case is $M = N = H = 2$, with $r \in \{0, 1, 2\}$. Since the exact local values would be known at every point of a singular optimal set, the table would also be a test case for the estimators of Problem 6.1.

Problem 6.3 (MAIS-O77). *For the loss $L(A, B) = \frac{1}{2}\|BA - \Phi\|_F^2$ with $H > r$:*

- (a) *Compute the local invariants $(\lambda(w), m(w))$ at every point w of the fiber F_Φ , i.e., solve Problem 6.2, and identify the minimal stratum.*
- (b) *Compute the two-sided invariants $(\lambda(w), m(w))$ at every point of each saddle set C_k , $0 \leq k < r$.*

Part (b) moves from the fiber, where L is minimal, to the saddle sets, where the two-sided definition is needed. It is the substance of the opposing-staircases conjecture in Chapter 7.

CHAPTER 7

Staircases

In Section 3.5, we saw that gradient flow on the deep linear network descends a staircase of saddle points in training time. On the Bayesian side, the posterior descends a ladder of free energies as the sample size grows. In this chapter, we compare the two staircases, which leads to the conjecture MAIS-O7 and the problems MAIS-O79 and MAIS-O81. The rungs of the dynamical staircase are saddle sets. MAIS-O81 asks whether a saddle has a free energy, and MAIS-O79 asks whether a saddle can be a local minimum of an averaged loss. Both lead to Laplace integrals over windows around saddles, which we compare with the Laplace and oscillatory integrals of Section 6.3.

Throughout this chapter, we use the deep linear network (3.1) with a target Φ of rank $r < H$ and distinct singular values $s_1 > \dots > s_r > 0$. Recall from Section 3.5 the critical sets C_k , which are saddle sets for $k < r$, and their losses

$$L_k = \frac{1}{2} \sum_{\alpha > k} s_\alpha^2 \quad \text{for } 0 \leq k \leq r.$$

7.1. The Bayesian ladder

Consider the Bayesian side of the same model, i.e., reduced rank regression with $M = N = H$. The free energy of a small region U_k around C_k is expected to be

$$F_n(U_k) \approx nS_n + nL_k + \lambda_k^{\text{AW}} \log n, \quad (7.1)$$

where λ_k^{AW} is the Aoyagi-Watanabe learning coefficient with true rank k (Section 6.6). The first term is common to all regions. The second term favors small loss, i.e., large k , and the third term favors small learning coefficients, i.e., small k . Hence, as n grows, the region with the smallest free energy, which carries most of the posterior mass, moves from $k = 0$ to larger k . This is the Bayesian ladder; see Figure 7.1.

Rung k overtakes rung $k - 1$ when

$$nL_{k-1} + \lambda_{k-1}^{\text{AW}} \log n = nL_k + \lambda_k^{\text{AW}} \log n.$$

Since $L_{k-1} - L_k = \frac{1}{2}s_k^2$, the crossover n_k is the large root of

$$\frac{n}{\log n} = \frac{2\Delta\lambda_k}{s_k^2}, \quad \text{where } \Delta\lambda_k = \lambda_k^{\text{AW}} - \lambda_{k-1}^{\text{AW}} > 0. \quad (7.2)$$

Suppose that the crossovers are ordered, i.e., $n_1 < n_2 < \dots < n_r$, which requires a well-separated spectrum. Then the posterior visits the ranks $0, 1, \dots, r$ in the same order in which gradient flow visits $C_0, C_1, \dots, C_r$.

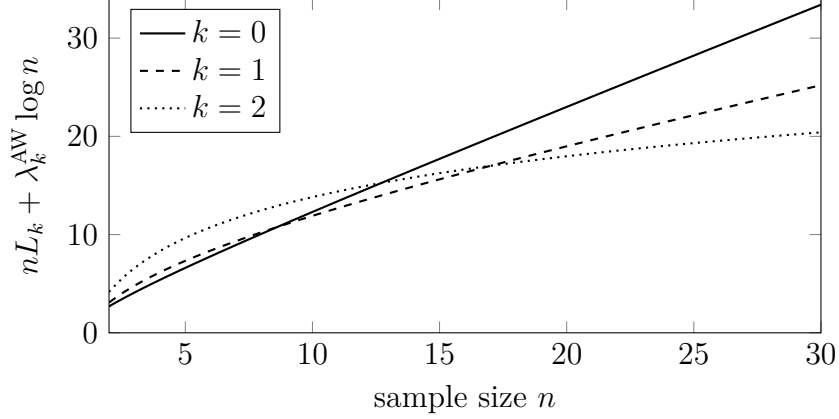


FIGURE 7.1. A schematic ladder with $(L_0, L_1, L_2) = (1, \frac{1}{2}, 0)$ and $(\lambda_0^{\text{AW}}, \lambda_1^{\text{AW}}, \lambda_2^{\text{AW}}) = (1, 3, 6)$: the lowest curve, which carries most of the posterior mass, moves from $k = 0$ to $k = 1$ and then to $k = 2$ as n grows

Remark. The approximation (7.1) is a theorem when the regions consist of minimizers at the same loss. It is conditional when they consist of local minimizers at different positive losses, since nL_n then fluctuates on the scale $\sqrt{n}$. For saddle sets, which are the rungs of the dynamical staircase, it is only a heuristic. We discuss both points in Section 7.3.

7.2. Opposing staircases

The learning coefficient is defined at local minima. To attach an invariant to the saddle sets C_k , we use the two-sided local learning coefficient of Chapter 6: the volume of nearby parameters whose loss lies within ε of $L(w)$, above or below, scales as $\varepsilon^{\lambda(w)}$ up to a logarithmic factor. For $0 \leq k \leq r$, let

$$\lambda_k = \inf_{w \in C_k} \lambda(w).$$

Conjecture (Opposing staircases, MAIS-O7). *For the two-layer linear network with $H > r$ and distinct singular values,*

$$\lambda_0 < \lambda_1 < \cdots < \lambda_r \quad \text{while} \quad L_0 > L_1 > \cdots > L_r.$$

That is, along the saddle chain visited by gradient flow, the loss falls and the learning coefficient rises, step for step. Since C_k is not compact, the attainment of the infimum is part of the claim.

If proved, the conjecture would let singular geometry predict the stages of training in advance, which is the central hope of developmental interpretability. Two pieces of evidence support it. In transformer experiments, the estimated local learning coefficient climbs as the loss falls through most training stages [HWFR]. On the Bayesian side, the Aoyagi-Watanabe coefficients of Section 6.6 have the positive increments that the conjecture demands. However, they are computed at the optimal sets of the models with true rank k , and not at the saddle sets. The computation of the two-sided values at the saddle sets is Problem 6.3(b).

The two-sided values at the saddle sets are small. Let $k < r$ and $(A, B) \in C_k$, and write $E = \Phi - BA$ as in Section 3.5. By the strict-saddle argument there, $L - L_k$ takes negative values in every neighborhood of (A, B) . It also takes positive values. Indeed, since $\text{rank } E = r - k \leq H$, we can write $E = -YX$ with $X \in M_{H,M}(\mathbb{R})$ and $Y \in M_{N,H}(\mathbb{R})$. Then

$$(B + tY)(A + tX) - \Phi = t(BX + YA) - (1 + t^2)E,$$

and $BX + YA$ is orthogonal to E in the Frobenius inner product, since $B^t E = 0$ and $EA^t = 0$. So

$$L(A + tX, B + tY) = (1 + t^2)^2 L_k + \frac{t^2}{2} \|BX + YA\|_F^2 > L_k \quad \text{for } t \neq 0.$$

Hence Lemma 6.5 gives

$$\lambda_k \leq 1 \quad \text{for } 0 \leq k < r.$$

On the other hand, $C_r = F_\Phi$ consists of minimizers. Suppose that $M = N = H$. Then $\lambda_r = \lambda_r^{\text{AW}}$, since the Aoyagi-Watanabe value is the smallest local learning coefficient on F_Φ , attained at its most degenerate points. Since λ_k^{AW} increases with k for $k < H$, and $H > r \geq 1$, the formula of Section 6.6 gives

$$\lambda_r \geq \lambda_0^{\text{AW}} = \frac{3H^2 + \delta_0}{8} \geq \frac{3}{2}.$$

So, for $M = N = H$, we have $\lambda_k \leq 1 < \lambda_k^{\text{AW}}$ for $k < r$, and the last inequality $\lambda_{r-1} < \lambda_r$ of the conjecture holds automatically. The rest of the conjecture concerns the r numbers $\lambda_0, \dots, \lambda_{r-1}$ in $(0, 1]$.

The conjecture as stated fails for some M and N . For $M = N = 1$, $H = 2$, and $r = 1$, write $A = a$ and $B = b^t$ with $a, b \in \mathbb{R}^2$. Then $C_0 = \{0\}$ and

$$L - L_0 = -s_1 \langle a, b \rangle + \frac{1}{2} \langle a, b \rangle^2.$$

So the origin is a non-degenerate critical point of index two in four variables, and $\lambda_0 = 1$ by the computation after Lemma 6.5. At every point of F_Φ , we have $\text{rank } A = 1 = M$, so $\lambda_1 = \frac{1}{2}$ by Section 6.6. Hence $\lambda_0 = 1 > \frac{1}{2} = \lambda_1$ in this case. So the conjecture needs an assumption on M and N , such as $M = N = H$, the case of the Bayesian ladder of Section 7.1.

7.3. The free energy of a saddle

The Bayesian ladder at saddle sets presupposes that a saddle point has a free energy of the form $nL + \Lambda \log n$ with a well-defined exponent Λ . The following example shows that this is not automatic.

In Laplace's method, an integral $\int_U e^{-nL}$ is governed by the points of $\bar{U}$ where L is smallest. For a window U around a local minimum, these are interior points, and Propositions 6.1 and 6.2 apply. For a window around a saddle point, they lie on the rim of the window, in the unstable direction.

Example (The rim of a Morse saddle). Let $L(x, y) = \frac{1}{2}(y^2 - x^2)$ on $\mathbb{R}^2$, with a Morse saddle at the origin, and let $U_\delta = [-\delta, \delta]^2$ with $\delta = \delta_n \rightarrow 0$ and $\delta_n \sqrt{n} \rightarrow \infty$. Then

$$\int_{U_\delta} e^{-nL(x,y)} dx dy = \int_{-\delta}^{\delta} e^{-\frac{n}{2}y^2} dy \int_{-\delta}^{\delta} e^{\frac{n}{2}x^2} dx = \sqrt{\frac{2\pi}{n}} \cdot \frac{2}{n\delta} e^{\frac{n}{2}\delta^2} (1 + o(1)).$$

The second integral is dominated by its endpoints. Indeed, the substitution $x = \delta - \frac{u}{n\delta}$ gives

$$\int_0^\delta e^{\frac{n}{2}x^2} dx = \frac{e^{\frac{n}{2}\delta^2}}{n\delta} \int_0^{n\delta^2} e^{-u(1 - \frac{u}{2n\delta^2})} du = \frac{e^{\frac{n}{2}\delta^2}}{n\delta} (1 + o(1)),$$

by dominated convergence, since the integrand is at most $e^{-\frac{u}{2}}$ and tends to e^{-u} . Since $\inf_{U_\delta} L = -\frac{1}{2}\delta^2$, the deterministic free energy

$$F(U_\delta) = -\log \int_{U_\delta} e^{-nL}$$

satisfies

$$F(U_\delta) - n \inf_{U_\delta} L = \frac{3}{2} \log n + \log \delta + O(1).$$

For $\delta = n^{-a}$ with $0 < a < \frac{1}{2}$, the coefficient of $\log n$ is $\frac{3}{2} - a$. So it depends on how the window shrinks, i.e., on the rim of the window in the escape direction, and not only on the saddle point.

In the language of Section 6.3, the minimum of L on U_δ is attained at the two points $(\pm\delta, 0)$, which lie on the edges $\{x = \pm\delta\}$, and

$$L - \inf_{U_\delta} L = \frac{1}{2}y^2 + \frac{1}{2}(\delta - |x|)(\delta + |x|).$$

Near these points, this function grows quadratically in y and linearly in $\delta - |x|$. So, for fixed δ , the volume function of $L - \inf_{U_\delta} L$ on U_δ satisfies

$$V(t) = \frac{8\sqrt{2}}{3\delta} t^{\frac{3}{2}} (1 + o(1)) \quad \text{as } t \rightarrow 0^+.$$

The exponent $\frac{3}{2}$ is $\frac{1}{2}$ from the stable direction plus 1 from the boundary in the unstable direction. Proposition 6.2 recovers the leading term above, since

$$\Gamma\left(\frac{5}{2}\right) \frac{8\sqrt{2}}{3\delta} = \frac{2\sqrt{2\pi}}{\delta}.$$

At the boundary scale $\delta = cn^{-\frac{1}{2}}$, the substitution $(x, y) = n^{-\frac{1}{2}}(X, Y)$ gives instead

$$F(U_\delta) - n \inf_{U_\delta} L = \log n + O(1).$$

The coefficient 1 is the two-sided learning coefficient of the saddle $x^2 - y^2$ in Chapter 6. However, the correction $-(m-1) \log \log n$ with $m = 2$, which the two-sided volume law would suggest, is absent. So particular windows reproduce particular invariants, and no window is distinguished.

Oscillatory integrals behave differently. For the same L and $\delta\sqrt{\tau} \rightarrow \infty$, integration by parts shows that each one-variable factor of $\int_{U_\delta} e^{i\tau L}$ differs from the Fresnel integral over $\mathbb{R}$ by $O(\frac{1}{\tau\delta})$. So

$$\int_{U_\delta} e^{i\tau L(x,y)} dx dy = \frac{2\pi}{\tau} \left(1 + O\left(\frac{1}{\delta\sqrt{\tau}}\right) \right).$$

Since $|e^{i\tau L}| = 1$, the rim only contributes lower order terms, and the decay rate τ^{-1} does not depend on the window. In the Laplace integral, the rim dominates by the factor $e^{\frac{n}{2}\delta^2}$. This is the difficulty in the following problem.

To state it, let $M = N = H$ and $1 \leq r < H$. Let the singular values $s_1(q) > \dots > s_r(q) > 0$ of the target $\Phi(q)$ depend on a parameter $q \rightarrow \infty$, such that

$$\frac{2\Delta\lambda_k}{s_k(q)^2} > e \quad \text{and} \quad \frac{2\Delta\lambda_k}{s_k(q)^2} \rightarrow \infty \quad \text{for } 1 \leq k \leq r.$$

Let $n_k(q)$ be the large root of (7.2) with $s_k = s_k(q)$. Suppose that the spectrum is well separated, i.e., $n_1(q) < \dots < n_r(q)$ for all large q . The data are independent pairs (x_j, y_j) , where x_j is standard Gaussian in $\mathbb{R}^H$ and $y_j = \Phi(q)x_j + \xi_j$ with standard Gaussian noise ξ_j . The prior φ is smooth and positive on a compact set $W \subset M_{H,H}(\mathbb{R})^2$ which meets every $C_k(q)$. Write

$$F_{q,n}(U) = -\log \int_U \varphi(w) \prod_{j=1}^n p(y_j | x_j, w) dw \quad \text{and} \quad \ell_{q,n}(w) = -\frac{1}{n} \sum_{j=1}^n \log p(y_j | x_j, w).$$

Let

$$J_q = \left[\sqrt{n_1(q)}, 2n_r(q) \right] \cap \mathbb{N}.$$

This range of sample sizes contains every crossover.

Problem 7.1 (MAIS-O81). *Either define a functional $\Lambda(C) \in \mathbb{Q}_{>0}$ on compact critical sets C of real-analytic losses, which equals the smallest local learning coefficient on C when C consists of local minimizers, and window radii $\delta_{q,n} > 0$ with $\sup_{n \in J_q} \delta_{q,n} \rightarrow 0$, such that for every $0 \leq k \leq r$, with $U_k^{q,n} = \{w \in W : \text{dist}(w, C_k(q) \cap W) < \delta_{q,n}\}$,*

$$\sup_{n \in J_q} \frac{|F_{q,n}(U_k^{q,n}) - n \inf_{U_k^{q,n}} \ell_{q,n} - \Lambda(C_k(q) \cap W) \log n|}{\log n} \rightarrow 0 \quad \text{in probability as } q \rightarrow \infty;$$

or prove that no such pair $(\Lambda, \delta_{q,n})$ exists, and identify the correction terms depending on the windows, such as the rim term above, that the ladder must carry.

Remark (Fluctuations at positive loss). In the notation of Chapter 6,

$$nL_n(w) - nL(w) = \sum_{j=1}^n \left(\log \frac{q_0(X_j)}{p(X_j | w)} - L(w) \right)$$

is a sum of n independent terms of mean zero and some variance $\sigma(w)^2$. On W_0 , the terms vanish. If $L(w) > 0$ and $p(\cdot | w)$ has the same support as q_0 , as for the Gaussian models below, then $\sigma(w)^2 > 0$, and the sum is of order $\sqrt{n}$, by the central limit theorem. For the model of Problem 7.1, we use the conditional densities $p(y | x, w)$ in place of $p(x | w)$, so that $\ell_{q,n} = S_n + L_n$. For $w = (A, B) \in C_k$, write, as in Section 3.5,

$$E = \Phi - BA = \sum_{\alpha > k} s_\alpha u_\alpha v_\alpha^t.$$

Then, with $y = \Phi x + \xi$,

$$\log \frac{q_0(y | x)}{p(y | x, w)} = \frac{1}{2} |Ex|^2 + \langle Ex, \xi \rangle.$$

Since x and ξ are independent standard Gaussian vectors,

$$\sigma(w)^2 = \sum_{\alpha > k} s_\alpha^2 + \frac{1}{2} \sum_{\alpha > k} s_\alpha^4.$$

For a fixed target and $k < r$, $\sqrt{n}\sigma(w)$ is much larger than $\lambda_k^{\text{AW}} \log n$. So, at positive losses, (7.1) can only hold after the random level of the loss on U_k is subtracted. This is why Problem 7.1 subtracts $n \inf \ell_{q,n}$ over the window, which contains both nS_n and the random level $n \inf L_n$.

For a fixed target, the crossovers n_k are fixed numbers, and no statement as $n \rightarrow \infty$ sees the ladder. So Problem 7.1 uses targets $\Phi(q)$ with $s_k(q) \rightarrow 0$, for which $n_k(q) \rightarrow \infty$. By (7.2),

$$n_k s_k^2 = 2\Delta \lambda_k \log n_k.$$

So, near the crossover, the difference $\frac{n}{2} s_k^2$ of the loss terms of the rungs $k-1$ and k is of the same order as the difference $\Delta \lambda_k \log n$ of their volume terms. The random difference between the two rungs is smaller. Indeed, on C_{k-1} , the matrix $\Phi - BA$ is $E + s_k u_k v_k^t$, and $E^t u_k = 0$. So the difference of the log-likelihood ratios above on C_{k-1} and on C_k is

$$s_k \langle v_k, x \rangle \langle u_k, \xi \rangle + \frac{1}{2} s_k^2 \langle v_k, x \rangle^2,$$

whose variance is $s_k^2 + \frac{1}{2} s_k^4$. Summed over the n samples, the random difference is of order

$$\sqrt{n} s_k \approx \sqrt{2\Delta \lambda_k \log n} = o(\log n).$$

This is why Problem 7.1 asks for the precision $o_p(\log n)$. Finally, the condition $\frac{2\Delta\lambda_k}{s_k(q)^2} > e$ guarantees that (7.2) has a large root, since $n \mapsto \frac{n}{\log n}$ decreases on $(1, e)$ and increases on (e, ∞) , with the minimum e .

Until Problem 7.1 is settled, the stratification by the learning coefficient is meaningful on the optimal set, but only heuristic on the saddle sets C_k with $k < r$, which gradient flow visits.

7.4. Effective losses

Researchers in developmental interpretability say that during a plateau, “the saddle is a local minimum of an effective loss”. Here is one way to make this precise.

Definition (Effective loss). For $\varepsilon, \delta > 0$, the effective loss at resolution (ε, δ) is

$$L_{\varepsilon, \delta}(w) = -\varepsilon \log \int_{B(w, \delta)} e^{-\frac{L(u)}{\varepsilon}} du.$$

A compact set C is an effective local minimum at resolution (ε, δ) if

$$\sup_{w \in C} L_{\varepsilon, \delta}(w) < \inf \{L_{\varepsilon, \delta}(w) : \text{dist}(w, C) = 2\delta\}.$$

So $e^{-\frac{1}{\varepsilon}L_{\varepsilon, \delta}(w)}$ is a Laplace integral over the ball $B(w, \delta)$, with the large parameter $\frac{1}{\varepsilon}$ in place of n . Hence $L_{\varepsilon, \delta}(w)$ is ε times the free energy of the window $B(w, \delta)$. Suppose that $L - \min_{\overline{B(w, \delta)}} L$ satisfies a volume law on $B(w, \delta)$ with the exponents (λ_w, m_w) . Then the proof of Proposition 6.2 gives

$$L_{\varepsilon, \delta}(w) = \min_{B(w, \delta)} L + \varepsilon \left(\lambda_w \log \frac{1}{\varepsilon} - (m_w - 1) \log \log \frac{1}{\varepsilon} + O(1) \right) \quad \text{as } \varepsilon \rightarrow 0.$$

The first term is the depth of the loss in the window, and the second term measures its volume. So the effective loss trades the depth of the loss in a window against its volume. At a degenerate saddle, a large volume may compensate for a small depth, so that the effective loss there may be smaller than at the points at the distance 2δ from it. As $\varepsilon \rightarrow 0$ with δ fixed, $L_{\varepsilon, \delta}(w) \rightarrow \min_{\overline{B(w, \delta)}} L$, uniformly on compact sets. So only the depth counts, and a compact set of strict saddles is not an effective minimum at small ε , at least when δ is small. For large ε , the average of L over the window dominates; see the remark below.

Remark (The effective loss as a free energy). Fix w , write $U = B(w, \delta)$, and let

$$\mu_w = \frac{e^{-\frac{L}{\varepsilon}}}{\int_U e^{-\frac{L}{\varepsilon}} du} \quad \text{on } U$$

be the Gibbs density of the window at the temperature ε . Proposition 1.4, with the uniform density on U as the prior, gives

$$L_{\varepsilon, \delta}(w) = \min_{\mu} \left(\int_U L \mu du + \varepsilon \int_U \mu \log \mu du \right),$$

where the minimum is over the probability densities μ on U , and it is attained at $\mu = \mu_w$. So $L_{\varepsilon, \delta}(w)$ is the mean energy minus the temperature times the entropy, minimized over the densities on the window, as in Section 1.5. In particular, the normalized indicator function of the sublevel set $\{u \in U : L(u) < t\}$ gives the bound of that section by the depth and the volume,

$$L_{\varepsilon, \delta}(w) \leq t - \varepsilon \log \text{Vol}(\{u \in U : L(u) < t\}) \quad \text{for } t > \min_{\overline{U}} L.$$

Differentiation under the integral sign in

$$L_{\varepsilon,\delta}(w) = -\varepsilon \log \int_{B(0,\delta)} e^{-\frac{1}{\varepsilon}L(w+v)} dv$$

gives

$$\begin{aligned} \nabla L_{\varepsilon,\delta}(w) &= \int_U \nabla L \mu_w du, \\ \nabla^2 L_{\varepsilon,\delta}(w) &= \int_U \nabla^2 L \mu_w du - \frac{1}{\varepsilon} \left(\int_U \nabla L \nabla L^t \mu_w du - \nabla L_{\varepsilon,\delta}(w) \nabla L_{\varepsilon,\delta}(w)^t \right). \end{aligned}$$

So the effective gradient flow in Problem 7.2(c) below follows the gradient of L averaged over the Gibbs density of the window, and the covariance of the gradient lowers the effective Hessian.

As $\varepsilon \rightarrow \infty$, the expansion of $e^{-\frac{L}{\varepsilon}}$ in powers of $\frac{1}{\varepsilon}$ gives

$$L_{\varepsilon,\delta}(w) = -\varepsilon \log \text{Vol}(U) + \langle L \rangle_{w,\delta} - \frac{1}{2\varepsilon} \text{Var}_{w,\delta}(L) + O(\varepsilon^{-2}),$$

where $\langle \cdot \rangle_{w,\delta}$ and $\text{Var}_{w,\delta}$ are the mean and the variance with respect to the normalized Lebesgue measure on U . The first term does not depend on w . So, at large ε , the effective loss is, up to a constant, the ball average of L , a mollification of L . The normalized Lebesgue measure on $B(0, \delta) \subset \mathbb{R}^d$ has vanishing odd moments, and

$$\langle v_i v_j \rangle_{0,\delta} = 0 \quad \text{for } i \neq j \quad \text{and} \quad \langle v_i^2 \rangle_{0,\delta} = \frac{\delta^2}{d+2}.$$

So Taylor's formula gives, with the Laplacian Δ on $\mathbb{R}^d$,

$$\langle L \rangle_{w,\delta} = L(w) + \frac{\delta^2}{2(d+2)} \Delta L(w) + O(\delta^4).$$

For the deep linear network, $d = H(M+N)$, and a direct computation gives

$$\Delta L = M \|B\|_F^2 + N \|A\|_F^2.$$

Since L is a polynomial of degree four, its fourth derivatives are constant, and the remainder $O(\delta^4)$ is a constant c which depends only on δ , M , N , and H . Hence

$$\langle L \rangle_{w,\delta} = L(w) + \frac{\delta^2}{2(d+2)} (M \|B\|_F^2 + N \|A\|_F^2) + c.$$

So, at large ε , the effective loss of the deep linear network is the loss plus a quadratic penalty on the weights. In machine learning, such a penalty is called weight decay.

A quadratic Morse saddle is never an effective local minimum. For $L(x, y) = \frac{1}{2}(y^2 - x^2)$, the translation by $(2\delta, 0)$ maps $B(0, \delta)$ onto $B((2\delta, 0), \delta)$, and for $|x| \leq \delta$,

$$L(x + 2\delta, y) - L(x, y) = -2\delta(x + \delta) \leq 0.$$

Hence $L_{\varepsilon,\delta}(2\delta, 0) < L_{\varepsilon,\delta}(0, 0)$ for all $\varepsilon, \delta > 0$, and $\{0\}$ is not an effective local minimum at any resolution. So the problem concerns degenerate saddles. In the deep linear network, the origin lies in C_0 , and the quadratic part of L at the origin is

$$-\text{Tr}(\Phi^t B A) = -\sum_{h=1}^H b_h^t \Phi a_h,$$

where a_h^t is the h -th row of A and b_h is the h -th column of B . Each summand is a quadratic form of rank $2r$ in $(a_h, b_h) \in \mathbb{R}^M \times \mathbb{R}^N$. So the Hessian of L at the origin has a kernel of dimension $H(M+N-2r)$.

Problem 7.2 (MAIS-O79). *Fix $R > 0$ so large that the ball $\{\|A\|_F^2 + \|B\|_F^2 \leq R\}$ meets every C_k , and let $C_k^R = C_k \cap \{\|A\|_F^2 + \|B\|_F^2 \leq R\}$.*

- (a) *For each k , determine the set $E_k \subset (0, \infty)^2$ of resolutions (ε, δ) at which C_k^R is an effective local minimum.*
- (b) *For $\delta > 0$, let $I_k(\delta) = \{\varepsilon > 0 : (\varepsilon, \delta) \in E_k\}$. Is there a δ for which every $I_k(\delta)$ is a nonempty interval and*

$$\sup I_k(\delta) < \inf I_{k-1}(\delta) \quad \text{for } 1 \leq k \leq r?$$

If so, decreasing ε moves the effective minima through $C_0^R, C_1^R, \dots, C_r^R$ in order.

- (c) *Fix such a δ . Let $w(t)$ be the gradient flow of L from the aligned initialization of Theorem 3.3 with the scale a_0 , and let w_{eff} solve $\dot{w}_{\text{eff}}(t) = -\nabla L_{\varepsilon_{a_0}(t), \delta}(w_{\text{eff}}(t))$ with the same initial point. Is there, for each small a_0 , a positive decreasing C^1 function ε_{a_0} and a time $T(a_0) \geq t_r$ such that*

$$\sup_{0 \leq t \leq T(a_0)} \text{dist}(w_{\text{eff}}(t), w(t)) \rightarrow 0 \quad \text{as } a_0 \rightarrow 0?$$

The truncation to C_k^R is needed, since the definition requires compact sets and C_k is not compact. Parts (a) and (b) are finite-dimensional questions about explicit Laplace integrals over balls near degenerate saddles, with the rim effects of Section 7.3. Part (c) asks whether annealing in ε is a proxy for training time; no principled choice of the schedule $\varepsilon(t)$ is known.

CHAPTER 8

Diffusions and metastability

Metastability is the phenomenon that a random process spends long periods near certain states, called metastable states, and only rarely moves between them. For the Langevin dynamics at small temperature ε , the metastable states are the wells of the potential, and the moves are the crossings of the barriers between them. In this chapter, we make the Langevin dynamics precise, relate it to partial differential equations, solve the one-dimensional case completely, and present the Eyring-Kramers formula, the potential-theoretic approach of Bovier-Eckhoff-Gayard-Klein [BEGK], and the semiclassical point of view. Our presentation mainly follows [BEGK] and Berglund [B].

Throughout this chapter, $L : \mathbb{T}^d \rightarrow [0, \infty)$ is a C^3 function, called the potential, $\varepsilon > 0$ is the temperature, and

$$\pi_\varepsilon(dw) = e^{-\frac{L(w)}{\varepsilon}} dw$$

is the (unnormalized) Gibbs measure. In the applications, L is a loss and w is a parameter.

8.1. Brownian motion and the Langevin dynamics

Definition (Brownian motion). A standard Brownian motion in $\mathbb{R}^d$ is a random continuous path $t \mapsto B_t$, $t \geq 0$, with $B_0 = 0$, such that for $0 \leq t_0 < t_1 < \dots < t_k$, the increments $B_{t_1} - B_{t_0}, \dots, B_{t_k} - B_{t_{k-1}}$ are independent Gaussian random vectors with mean 0 and covariance matrices $(t_1 - t_0)\mathbf{I}, \dots, (t_k - t_{k-1})\mathbf{I}$.

The existence of Brownian motion was proved by Wienerⁱ.

The Langevin dynamics with the initial point w is the shorthand for the integral equation

$$w_t = w - \int_0^t \nabla L(w_s) ds + \sqrt{2\varepsilon} B_t \quad \text{for } t \geq 0, \quad (8.1)$$

where ∇L is regarded as a periodic vector field on $\mathbb{R}^d$, and w_t is projected to $\mathbb{T}^d$. Since the noise is additive and ∇L is Lipschitz, (8.1) can be solved path by path, by the Picard iteration as for ordinary differential equations. We write $\mathbb{P}_w$ and $\mathbb{E}_w$ for the probability and the expectation for the process started at w . For $\varepsilon = 0$, (8.1) is the gradient flow, along which L decreases. The noise allows the process to climb up the potential. For small ε , it does so only rarely.

The following formula is due to Itôⁱⁱ.

Theorem (Itô's formula). For $f \in C^2(\mathbb{T}^d)$,

$$f(w_t) = f(w) + \int_0^t \mathcal{L}_\varepsilon f(w_s) ds + \sqrt{2\varepsilon} \int_0^t \nabla f(w_s) \cdot dB_s, \quad \text{where } \mathcal{L}_\varepsilon = \varepsilon \Delta - \nabla L \cdot \nabla.$$

The last term is a stochastic integral, which has mean zero.

ⁱN. Wiener, *Differential-space*. J. Math. and Phys. 2 (1923), 131–174.

ⁱⁱK. Itô, *On a formula concerning stochastic differentials*. Nagoya Math. J. 3 (1951), 55–65.

Remark (The idea of Itô's formula). Suppose that $f \in C^3(\mathbb{T}^d)$. Let $0 = t_0 < t_1 < \dots < t_N = t$, and write

$$\delta t_k = t_{k+1} - t_k, \quad \delta w_k = w_{t_{k+1}} - w_{t_k}, \quad \text{and} \quad \delta B_k = B_{t_{k+1}} - B_{t_k}.$$

By Taylor's formula,

$$f(w_t) - f(w) = \sum_k \nabla f(w_{t_k}) \cdot \delta w_k + \frac{1}{2} \sum_k \delta w_k \cdot \nabla^2 f(w_{t_k}) \delta w_k + \sum_k O(|\delta w_k|^3).$$

By (8.1),

$$\delta w_k = -\nabla L(w_{t_k}) \delta t_k + \sqrt{2\varepsilon} \delta B_k + o(\delta t_k),$$

and δB_k is Gaussian with covariance $\delta t_k \mathbf{I}$. So $|\delta w_k|$ is of size $\delta t_k^{\frac{1}{2}}$, not δt_k , and the second-order sum does not vanish as $\max_k \delta t_k \rightarrow 0$. Its main part is

$$\varepsilon \sum_k \delta B_k \cdot \nabla^2 f(w_{t_k}) \delta B_k.$$

Given the path up to time t_k , the k -th term has the mean and the variance

$$\varepsilon \Delta f(w_{t_k}) \delta t_k \quad \text{and} \quad 2\varepsilon^2 \text{Tr}(\nabla^2 f(w_{t_k})^2) \delta t_k^2,$$

and the fluctuations of different terms are uncorrelated. So the second-order sum tends to $\varepsilon \int_0^t \Delta f(w_s) ds$ in L^2 . This is the rule $dB_j dB_l = \delta_{jl} dt$. The third-order sum is of size $\sum_k \delta t_k^{\frac{3}{2}}$, which tends to 0. The first-order sum tends to $-\int_0^t \nabla L \cdot \nabla f(w_s) ds$ plus the limit of

$$\sqrt{2\varepsilon} \sum_k \nabla f(w_{t_k}) \cdot \delta B_k,$$

which is, by definition, the stochastic integral. Each term has mean zero, since $\nabla f(w_{t_k})$ depends only on the path up to time t_k , and δB_k is independent of it and has mean zero. This is why the integrand is evaluated at the left endpoints.

The operator $\mathcal{L}_\varepsilon$ is called the generator of the Langevin dynamics. Taking expectations, we obtain

$$\mathbb{E}_w[f(w_t)] = f(w) + t\mathcal{L}_\varepsilon f(w) + o(t) \quad \text{as } t \rightarrow 0.$$

The term $\varepsilon \Delta$ comes from the variance $2\varepsilon t$ of the noise, and the term $-\nabla L \cdot \nabla$ comes from the drift. Moreover, $u(t, w) = \mathbb{E}_w[f(w_t)]$ solves the Kolmogorovⁱ equation

$$\partial_t u = \mathcal{L}_\varepsilon u \quad \text{with } u(0, \cdot) = f.$$

See also Delande [D, Section 1.1].

Proposition 8.1. For $f, g \in C^2(\mathbb{T}^d)$,

$$\int_{\mathbb{T}^d} f \mathcal{L}_\varepsilon g d\pi_\varepsilon = -\varepsilon \int_{\mathbb{T}^d} \nabla f \cdot \nabla g d\pi_\varepsilon.$$

Consequently, the Gibbs measure is invariant, that is,

$$\int_{\mathbb{T}^d} \mathbb{E}_w[g(w_t)] \pi_\varepsilon(dw) = \int_{\mathbb{T}^d} g d\pi_\varepsilon \quad \text{for } t \geq 0.$$

Moreover, $-\mathcal{L}_\varepsilon$ is a symmetric non-negative operator on $L^2(\mathbb{T}^d, \pi_\varepsilon)$.

ⁱA. Kolmogoroff, *Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung*. Math. Ann. 104 (1931), 415–458.

PROOF. Since

$$\nabla e^{-\frac{L}{\varepsilon}} = -\frac{1}{\varepsilon} e^{-\frac{L}{\varepsilon}} \nabla L,$$

we have

$$e^{-\frac{L}{\varepsilon}} \mathcal{L}_\varepsilon g = \varepsilon e^{-\frac{L}{\varepsilon}} \Delta g - e^{-\frac{L}{\varepsilon}} \nabla L \cdot \nabla g = \varepsilon \operatorname{div} \left(e^{-\frac{L}{\varepsilon}} \nabla g \right).$$

Multiplying by f and integrating by parts on $\mathbb{T}^d$ gives the identity. Taking $f = 1$, we get

$$\int_{\mathbb{T}^d} \mathcal{L}_\varepsilon g \, d\pi_\varepsilon = 0.$$

So, by the Kolmogorov equation for $u(t, w) = \mathbb{E}_w[g(w_t)]$,

$$\frac{d}{dt} \int_{\mathbb{T}^d} \mathbb{E}_w[g(w_t)] \pi_\varepsilon(dw) = \int_{\mathbb{T}^d} \mathcal{L}_\varepsilon u(t, \cdot) \, d\pi_\varepsilon = 0.$$

The last claim follows from the symmetry of the right-hand side in (f, g) , and from

$$-\int_{\mathbb{T}^d} f \mathcal{L}_\varepsilon f \, d\pi_\varepsilon = \varepsilon \int_{\mathbb{T}^d} |\nabla f|^2 \, d\pi_\varepsilon \geq 0.$$

□

The symmetry is the reversibility of the dynamics with respect to π_ε ; see Berglund [B, Section 1]. As $\varepsilon \rightarrow 0$, the Gibbs measure concentrates on the zero set $L^{-1}(0)$, and the process spends most of its time near it.

In the language of semigroups, the Kolmogorov equation says that

$$\mathbb{E}_w[f(w_t)] = (e^{t\mathcal{L}_\varepsilon} f)(w) \quad \text{for } t \geq 0.$$

So the Langevin dynamics is encoded in the semigroup $e^{t\mathcal{L}_\varepsilon}$, a heat semigroup with a drift. The Schwartz kernel $p_t(w, w')$ of $e^{t\mathcal{L}_\varepsilon}$, defined by

$$(e^{t\mathcal{L}_\varepsilon} f)(w) = \int_{\mathbb{T}^d} p_t(w, w') f(w') \, dw',$$

is the transition density: $p_t(w, \cdot)$ is the density of the law of w_t under $\mathbb{P}_w$. As a function of (t, w') , it solves the adjoint equation with respect to dw , called the Fokker-Planck equation,

$$\partial_t \rho = \varepsilon \Delta \rho + \operatorname{div}(\rho \nabla L) = \varepsilon \operatorname{div} \left(e^{-\frac{L}{\varepsilon}} \nabla \left(e^{\frac{L}{\varepsilon}} \rho \right) \right) \quad \text{with } \rho(0, \cdot) = \delta_w.$$

By the identity in the proof of Proposition 8.1, $\rho = g e^{-\frac{L}{\varepsilon}}$ solves the Fokker-Planck equation if and only if $\partial_t g = \mathcal{L}_\varepsilon g$. In particular, the stationary solutions are the multiples of $e^{-\frac{L}{\varepsilon}}$, which is the invariance of the Gibbs measure written for densities. On $\mathbb{R}^d$ instead of $\mathbb{T}^d$, two examples are explicit.

(1) *No potential*: If L is constant, then $w_t = w + \sqrt{2\varepsilon} B_t$, and

$$p_t(w, w') = (4\pi\varepsilon t)^{-\frac{d}{2}} e^{-\frac{|w-w'|^2}{4\varepsilon t}}$$

is the heat kernel of $e^{t\varepsilon\Delta}$. So Brownian motion is the probabilistic form of the heat kernel.

(2) *Quadratic potential*: If $L(w) = \frac{\mu}{2}|w|^2$ with $\mu > 0$, then the law of w_t under $\mathbb{P}_w$ is Gaussian with mean $e^{-\mu t} w$ and covariance $\frac{\varepsilon}{\mu}(1 - e^{-2\mu t})\mathbf{I}$. As $t \rightarrow \infty$, it tends to the normalized Gibbs measure, the Gaussian with covariance $\frac{\varepsilon}{\mu}\mathbf{I}$. This is the Ornstein-Uhlenbeck process, and p_t is given by Mehler's formula.

In general, Lemma 8.4 gives

$$e^{t\mathcal{L}_\varepsilon} = e^{\frac{L}{2\varepsilon}} e^{-\frac{t}{\varepsilon} P_\varepsilon} e^{-\frac{L}{2\varepsilon}}.$$

So every expectation for the Langevin dynamics is computed by the Schrödinger semigroup of the Witten Laplacian P_ε . For the quadratic potential, P_ε is a harmonic oscillator. Finally, the reversibility means that if w_0 is distributed according to the normalized Gibbs measure, then the pair (w_0, w_t) has the same law as (w_t, w_0) . Indeed, by the symmetry in Proposition 8.1, $e^{-\frac{L(w)}{\varepsilon}} p_t(w, w')$ is symmetric in (w, w') .

8.2. Hitting times and boundary value problems

Definition (Hitting times). For a closed set $C \subset \mathbb{T}^d$, the hitting time of C is

$$\tau_C = \inf \{t > 0 : w_t \in C\}.$$

A basic tool is the *strong Markov property*: at a hitting time τ_C , the process starts afresh, i.e., given the path up to time τ_C , the future path $(w_{\tau_C+t})_{t \geq 0}$ is the Langevin dynamics started at w_{τ_C} .

Probabilities and expectations of hitting times solve boundary value problems for $\mathcal{L}_\varepsilon$. Let $A, B \subset \mathbb{T}^d$ be disjoint closed sets whose complements have C^1 boundaries. Then the probability $q_{A,B}(w) = \mathbb{P}_w[\tau_A < \tau_B]$ to reach A before B solves

$$\mathcal{L}_\varepsilon q_{A,B} = 0 \text{ in } \mathbb{T}^d \setminus (A \cup B), \quad q_{A,B} = 1 \text{ on } A, \quad q_{A,B} = 0 \text{ on } B,$$

and the mean hitting time $w_B(w) = \mathbb{E}_w[\tau_B]$ solves

$$-\mathcal{L}_\varepsilon w_B = 1 \text{ in } \mathbb{T}^d \setminus B, \quad w_B = 0 \text{ on } B;$$

see [BEGK, (2.9), (2.17), and (2.25)]. The function $q_{A,B}$ is called the equilibrium potential, or the committor function [B, Section 3].

A random time $\tau \geq 0$ is called a stopping time if, for each $t \geq 0$, the event $\{\tau \leq t\}$ is determined by the path $(w_s)_{0 \leq s \leq t}$. Hitting times are stopping times; the time of the last visit to a set before time 1 is not. The stochastic integral in Itô's formula, stopped at a bounded stopping time, still has mean zero. So, for $f \in C^2(\mathbb{T}^d)$, a stopping time τ , and $t \geq 0$, taking expectations in Itô's formula at the time $t \wedge \tau = \min\{t, \tau\}$ gives Dynkin's formula

$$\mathbb{E}_w[f(w_{t \wedge \tau})] = f(w) + \mathbb{E}_w \left[\int_0^{t \wedge \tau} \mathcal{L}_\varepsilon f(w_s) ds \right].$$

Dynkin's formula proves the representations above, starting from the solutions of the boundary value problems, which exist by elliptic theory. Suppose that A and B have smooth boundaries, and let u solve

$$-\mathcal{L}_\varepsilon u = 1 \text{ in } \mathbb{T}^d \setminus B, \quad u = 0 \text{ on } \partial B.$$

By elliptic regularity, u is C^2 up to ∂B , and we extend it to a function in $C^2(\mathbb{T}^d)$. By the maximum principle, $u \geq 0$ outside B . Let $w \notin B$. For $s < \tau_B$, we have $w_s \notin B$, and hence $\mathcal{L}_\varepsilon u(w_s) = -1$. So Dynkin's formula with $\tau = \tau_B$ gives

$$u(w) = \mathbb{E}_w[u(w_{t \wedge \tau_B})] + \mathbb{E}_w[t \wedge \tau_B].$$

Hence $\mathbb{E}_w[t \wedge \tau_B] \leq \max u$ for all t . By monotone convergence, $\mathbb{E}_w[\tau_B] \leq \max u < \infty$, and in particular, $\tau_B < \infty$ almost surely. As $t \rightarrow \infty$, $u(w_{t \wedge \tau_B}) \rightarrow u(w_{\tau_B}) = 0$, and dominated convergence gives

$$u(w) = \mathbb{E}_w[\tau_B].$$

The same argument for the solution q of $\mathcal{L}_\varepsilon q = 0$ in $\mathbb{T}^d \setminus (A \cup B)$, with $q = 1$ on ∂A and $q = 0$ on ∂B , gives

$$q(w) = \mathbb{E}_w [q(w_{\tau_{A \cup B}})] = \mathbb{P}_w [\tau_A < \tau_B].$$

So the solutions of the two boundary value problems are the equilibrium potential $q_{A,B}$ and the mean hitting time w_B .

8.3. Metastability in dimension one

In dimension one, the mean hitting time can be computed explicitly, and the whole asymptotic theory is an exercise in Laplace's method. We consider the Langevin dynamics on $\mathbb{R}$ with a potential $L \in C^3(\mathbb{R})$ which grows at least linearly at $-\infty$, i.e., $L(z) \geq |z|$ for $z \leq -R$, for some $R > 0$.

Proposition 8.2. *For $b \in \mathbb{R}$, the function*

$$w(x) = \frac{1}{\varepsilon} \int_x^b e^{\frac{L(y)}{\varepsilon}} \left(\int_{-\infty}^y e^{-\frac{L(z)}{\varepsilon}} dz \right) dy \quad \text{for } x \leq b$$

solves

$$\varepsilon w'' - L'w' = -1 \quad \text{on } (-\infty, b) \quad \text{with } w(b) = 0.$$

Moreover, $e^{-\frac{L}{\varepsilon}} w' \rightarrow 0$ at $-\infty$.

PROOF. Since

$$\varepsilon w'' - L'w' = \varepsilon e^{\frac{L}{\varepsilon}} \left(e^{-\frac{L}{\varepsilon}} w' \right)',$$

the equation is equivalent to

$$\left(e^{-\frac{L}{\varepsilon}} w' \right)' = -\frac{1}{\varepsilon} e^{-\frac{L}{\varepsilon}}.$$

Differentiating the formula gives

$$e^{-\frac{L(x)}{\varepsilon}} w'(x) = -\frac{1}{\varepsilon} \int_{-\infty}^x e^{-\frac{L(z)}{\varepsilon}} dz,$$

which satisfies this equation, and which tends to 0 at $-\infty$ by the growth of L . $\square$

For $a < x < b$, Dynkin's formula on the interval (a, b) , as in the previous section, gives

$$w(x) = w(a) \mathbb{P}_x [\tau_a < \tau_b] + \mathbb{E}_x [\tau_a \wedge \tau_b].$$

One checks that the first term tends to 0 as $a \rightarrow -\infty$, by the growth of L , and the second term tends to $\mathbb{E}_x [\tau_b]$ by monotone convergence. So $w(x) = \mathbb{E}_x [\tau_b]$ is the mean hitting time of b . The next theorem is the one-dimensional Eyring-Kramers formula, in which we allow the well to be degenerate; see Figure 8.1.

Theorem 8.3. *Let $x_A < z^* < b$, and assume the following.*

- (i) x_A is the unique global minimum of L on $(-\infty, b]$, and there are $\kappa > 0$ and an integer $k \geq 1$ such that

$$L(z) - L(x_A) = \kappa (z - x_A)^{2k} (1 + O(|z - x_A|)) \quad \text{as } z \rightarrow x_A.$$

- (ii) $L'(z^*) = 0$, $L''(z^*) = -\mu_- < 0$, and $L(y) < L(z^*)$ for $y \in (z^*, b]$.

Let $h = L(z^*) - L(x_A)$. Let K be a compact subset of $\{x < z^* : L(y) < L(z^*) \text{ for } y \in [x, z^*)\}$. Then, uniformly for $x \in K$,

$$\mathbb{E}_x [\tau_b] = 2\Gamma \left(1 + \frac{1}{2k} \right) \kappa^{-\frac{1}{2k}} \sqrt{\frac{2\pi}{\mu_-}} \varepsilon^{\frac{1}{2k} - \frac{1}{2}} e^{\frac{h}{\varepsilon}} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0.$$

PROOF. Write

$$I(y) = \int_{-\infty}^y e^{-\frac{L(z)}{\varepsilon}} dz.$$

By Proposition 8.2,

$$\mathbb{E}_x[\tau_b] = \frac{1}{\varepsilon} \int_x^b e^{\frac{L(y)}{\varepsilon}} I(y) dy.$$

We first estimate the inner integral. Let

$$J = 2\Gamma\left(1 + \frac{1}{2k}\right) \left(\frac{\varepsilon}{\kappa}\right)^{\frac{1}{2k}} e^{-\frac{L(x_A)}{\varepsilon}}.$$

We claim that $I(y) = J(1 + o(1))$ uniformly for $y \in [\frac{x_A + z^*}{2}, b]$. Let $\theta \in (0, 1)$. Choose $\delta \in (0, \frac{z^* - x_A}{2})$ such that

$$(1 - \theta)\kappa(z - x_A)^{2k} \leq L(z) - L(x_A) \leq (1 + \theta)\kappa(z - x_A)^{2k} \quad \text{for } |z - x_A| \leq \delta.$$

We use the substitution $z - x_A = (\frac{\varepsilon}{\kappa})^{\frac{1}{2k}} v$ and the integral

$$\int_{\mathbb{R}} e^{-cv^{2k}} dv = 2\Gamma\left(1 + \frac{1}{2k}\right) c^{-\frac{1}{2k}} \quad \text{for } c > 0.$$

So

$$(1 + \theta)^{-\frac{1}{2k}} J(1 - o(1)) \leq \int_{|z - x_A| \leq \delta} e^{-\frac{L(z)}{\varepsilon}} dz \leq (1 - \theta)^{-\frac{1}{2k}} J.$$

On the rest of $(-\infty, b]$, we have $L \geq L(x_A) + c$ for some $c > 0$, by (i) and the growth of L . So the rest contributes

$$O\left(e^{-\frac{L(x_A) + c}{\varepsilon}}\right) = o(J).$$

Letting $\theta \rightarrow 0$ proves the claim.

Next, we estimate the outer integral. Let $\delta > 0$ be small. For $x \in K$ and $y \in [x, b]$ with $|y - z^*| \geq \delta$, we have $L(y) \leq L(z^*) - c$ for some $c > 0$ independent of $x \in K$, by the choice of K , assumption (ii), and compactness. Since $I(y) \leq I(b) = J(1 + o(1))$, the part of the outer integral with $|y - z^*| \geq \delta$ is

$$O\left(e^{\frac{L(z^*) - c}{\varepsilon}} J\right).$$

On $|y - z^*| \leq \delta$, the claim gives $I(y) = J(1 + o(1))$. Laplace's method at the non-degenerate maximum z^* , as in the proof of Proposition 6.1, gives

$$\int_{z^* - \delta}^{z^* + \delta} e^{\frac{L(y)}{\varepsilon}} dy = e^{\frac{L(z^*)}{\varepsilon}} \sqrt{\frac{2\pi\varepsilon}{\mu_-}} (1 + o(1)).$$

Hence

$$\mathbb{E}_x[\tau_b] = \frac{1}{\varepsilon} e^{\frac{L(z^*)}{\varepsilon}} \sqrt{\frac{2\pi\varepsilon}{\mu_-}} J(1 + o(1)),$$

which is the claimed formula. $\square$

Remark. For $k = 1$, we have

$$\kappa = \frac{1}{2} L''(x_A) \quad \text{and} \quad 2\Gamma\left(\frac{3}{2}\right) \kappa^{-\frac{1}{2}} = \sqrt{\frac{2\pi}{L''(x_A)}}.$$

So the formula becomes the classical Kramers formula

$$\mathbb{E}_x[\tau_b] = \frac{2\pi}{\sqrt{\mu_- L''(x_A)}} e^{\frac{h}{\varepsilon}} (1 + o(1)).$$

The degenerate well $\kappa(z - x_A)^{2k}$ has the learning coefficient $\lambda = \frac{1}{2k}$ (see the examples in Section 6.3), and the dimension is $d = 1$. So the transition time carries the factor $\varepsilon^{\lambda - \frac{d}{2}}$, and it is longer for a flat well than for a quadratic one by a power of ε . In dimension one, this is the law conjectured in MAIS-O74 (Chapter 9), and it holds for each starting point. In higher dimensions, there is no explicit formula, and the zero set of L can be positive-dimensional and singular.

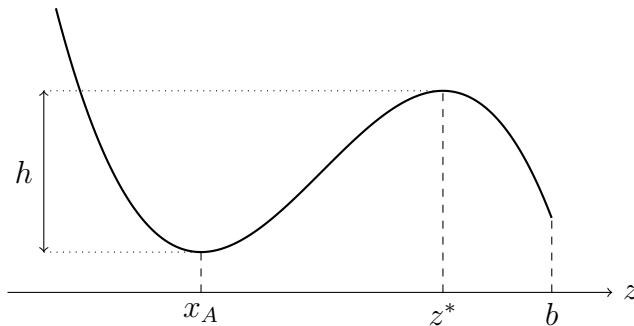


FIGURE 8.1. The potential $L(z) = -\frac{z^3}{3} + z + 1$, with a well at $x_A = -1$ and a barrier of height $h = \frac{4}{3}$ at $z^* = 1$

8.4. The Eyring-Kramers formula

In higher dimensions, the barrier between two wells is crossed near a saddle point.

Definition (Critical points and Morse saddles). A point z is a critical point of L if $\nabla L(z) = 0$. It is non-degenerate if the Hessian $\nabla^2 L(z)$ is invertible. Its index is the number of negative eigenvalues of $\nabla^2 L(z)$. A function whose critical points are all non-degenerate is called a Morse function. A non-degenerate critical point of index one is called a Morse saddle: L decreases in one direction and increases in the orthogonal directions.

The communication height between two disjoint compact sets K_A and K_B is

$$h = \inf_{\gamma} \max_{t \in [0,1]} L(\gamma(t)),$$

where the infimum is over the continuous paths $\gamma : [0, 1] \rightarrow \mathbb{T}^d$ from K_A to K_B . Generically, it is attained at a Morse saddle [B, Section 1]. On the exponential scale, the large deviation theory of Freidlin-Wentzellⁱ gives the Arrheniusⁱⁱ law

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \mathbb{E}_{x_A} [\tau_B] = L(z^*) - L(x_A)$$

for two non-degenerate minima x_A, x_B separated by a Morse saddle z^* , where τ_B is the hitting time of a small ball around x_B . Moreover, the exit time from a neighborhood of a stable equilibrium, divided by its mean, is asymptotically exponentially distributed [B, Section 2]. So the transitions are rare events occurring at exponential times. The Eyring-Kramers formula identifies the prefactor.

Remark (Hitting times and the Dirichlet heat semigroup). Let B be a closed set with smooth boundary such that $D = \mathbb{T}^d \setminus B$ is connected. Let $-\mathcal{L}_\varepsilon^D$ be the self-adjoint operator on $L^2(D, \pi_\varepsilon)$ defined by the quadratic form $\varepsilon \int_D |\nabla f|^2 d\pi_\varepsilon$ on $H_0^1(D)$, i.e., $-\mathcal{L}_\varepsilon$ with the Dirichlet condition

ⁱM. Freidlin and A. Wentzell, *Random Perturbations of Dynamical Systems*. Springer, 1984.

ⁱⁱS. Arrhenius, *Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren*. Z. Phys. Chem. 4 (1889), 226–248.

on ∂B . Let $0 < \Lambda_1 < \Lambda_2 \leq \Lambda_3 \leq \dots$ be its eigenvalues, with eigenfunctions ψ_j orthonormal in $L^2(D, \pi_\varepsilon)$. The survival probability $u(t, w) = \mathbb{P}_w[\tau_B > t]$ solves the Kolmogorov equation with the Dirichlet condition

$$\partial_t u = \mathcal{L}_\varepsilon u \text{ in } D, \quad u = 0 \text{ on } \partial B, \quad u(0, \cdot) = 1.$$

Hence

$$\mathbb{P}_w[\tau_B > t] = \left(e^{t\mathcal{L}_\varepsilon^D} 1 \right)(w) = \sum_j e^{-\Lambda_j t} \psi_j(w) \int_D \psi_j d\pi_\varepsilon.$$

Integrating in t , we obtain

$$\mathbb{E}_w[\tau_B] = \int_0^\infty \mathbb{P}_w[\tau_B > t] dt = \left((-\mathcal{L}_\varepsilon^D)^{-1} 1 \right)(w),$$

which is the boundary value problem for w_B . In the metastable situation, L is a Morse function, B is a small ball around a local minimum x_B , and x_A is the only local minimum of L in D . Then Λ_1 is exponentially small, while $\Lambda_2 \geq c > 0$ for small ε . The lower bound on Λ_2 follows from the harmonic approximation of Section 8.6, which applies also with the Dirichlet condition, since $\nabla L \neq 0$ on ∂B . Heuristically, ψ_1 is close to a constant c_ε on the well of x_A and small on the basin of x_B . So the measures $\psi_1 \pi_\varepsilon$ and $\psi_1^2 \pi_\varepsilon$ concentrate on the well of x_A , even if x_B is deeper and π_ε itself concentrates near ∂B . Hence

$$c_\varepsilon \int_D \psi_1 d\pi_\varepsilon \approx \int_D \psi_1^2 d\pi_\varepsilon = 1.$$

So, for w in the well and $t \gg 1$,

$$\mathbb{P}_w[\tau_B > t] \approx e^{-\Lambda_1 t} \quad \text{and} \quad \mathbb{E}_w[\tau_B] \approx \frac{1}{\Lambda_1}.$$

This is the exponential law of the transition times. To leading order, the Eyring-Kramers formula is the asymptotics of Λ_1^{-1} , the inverse of the principal Dirichlet eigenvalue, and metastability is the separation $\Lambda_1 \ll \Lambda_2$.

The bound $\Lambda_2 \geq c$ uses that the critical points of L are non-degenerate. At a degenerate well, Λ_2 can tend to 0, and it is expected to be only polynomially small in ε . For instance, let $d = 1$, and suppose that $L(z) = L(x_A) + \kappa(z - x_A)^{2k}$ near x_A with $k \geq 2$. Then the substitution $z - x_A = \varepsilon^{\frac{1}{2k}} y$ turns the Witten Laplacian P_ε of Lemma 8.4 near x_A into $\varepsilon^{2-\frac{1}{k}}$ times the operator

$$-\partial_y^2 + k^2 \kappa^2 y^{4k-2} - k(2k-1)\kappa y^{2k-2},$$

which does not depend on ε . Since $-\mathcal{L}_\varepsilon^D$ is unitarily equivalent to $\frac{1}{\varepsilon} P_\varepsilon$ with the Dirichlet condition, Λ_2 is of order $\varepsilon^{1-\frac{1}{k}}$. For the singular wells of Chapter 9, a polynomial lower bound on Λ_2 is open. It is closely related to the spectral route to Problem 9.2 through the dynamics reflected near the well; see the remark on spectral gap and relaxation in Section 8.6.

Theorem (Eyring-Kramers formula; Bovier-Eckhoff-Gayrard-Klein). *Suppose that L is a Morse function, x_A and x_B are local minima, and the communication height between them is attained at a single Morse saddle z^* , at which $\nabla^2 L(z^*)$ has the negative eigenvalue $-\mu_-$. Under the non-degeneracy assumptions of [BEGK],*

$$\mathbb{E}_{x_A}[\tau_B] = \frac{2\pi}{\mu_-} \sqrt{\frac{|\det \nabla^2 L(z^*)|}{\det \nabla^2 L(x_A)}} e^{\frac{L(z^*) - L(x_A)}{\varepsilon}} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0.$$

The formula goes back to Eyringⁱ and Kramersⁱⁱ in the 1930s and 1940s. The first rigorous proof in dimensions $d \geq 2$ is [BEGK, Theorem 3.2]. See Berglund [B, Section 1] for the history and the other approaches.

Remark (The exponential scale as an action). The theory of Freidlin-Wentzell is Laplace's method on the space of paths. Formally, the white noise $\dot{B}$ has the density

$$\exp\left(-\frac{1}{2}\int_0^T |\dot{B}_t|^2 dt\right),$$

since the increments of B over intervals of length δt are independent Gaussians with covariance $\delta t \mathbf{I}$. Solving (8.1) for $\dot{B}$, we see that the probability that $(w_t)_{0 \leq t \leq T}$ stays close to a given C^1 path ϕ is, formally, $e^{-\frac{S_T(\phi) + o(1)}{\varepsilon}}$, where

$$S_T(\phi) = \frac{1}{4} \int_0^T |\dot{\phi} + \nabla L(\phi)|^2 dt.$$

The action vanishes exactly along the gradient flow. For a path from x_A to z^* , since

$$|\dot{\phi} + \nabla L(\phi)|^2 = |\dot{\phi} - \nabla L(\phi)|^2 + 4\dot{\phi} \cdot \nabla L(\phi),$$

we have

$$S_T(\phi) \geq \int_0^T \dot{\phi} \cdot \nabla L(\phi) dt = L(z^*) - L(x_A).$$

Equality is approached as $T \rightarrow \infty$ along the reversed gradient flow $\dot{\phi} = \nabla L(\phi)$, i.e., along the branch of the unstable manifold of z^* which descends to x_A , run backward. So the cheapest way to climb is to run the gradient flow backward, at the cost of the height climbed. After z^* , the path descends to x_B at no cost. This is the Arrhenius law.

In the language of Section 8.6, the potential of P_ε is $V = \frac{1}{4}|\nabla L|^2$ to leading order. The Agmon distance is

$$d_{\text{Ag}}(x, y) = \inf_{\gamma} \int_0^1 \sqrt{V(\gamma)} |\dot{\gamma}| dt,$$

where the infimum is over the paths γ from x to y . For every path,

$$\int_0^1 \sqrt{V(\gamma)} |\dot{\gamma}| dt = \frac{1}{2} \int_0^1 |\nabla L(\gamma)| |\dot{\gamma}| dt \geq \frac{1}{2} |L(\gamma(1)) - L(\gamma(0))|,$$

with equality along gradient lines. So

$$d_{\text{Ag}}(x_A, x) = \frac{1}{2} (L(x) - L(x_A))$$

for x in the basin of x_A and for $x = z^*$. Hence the ground state $e^{-\frac{L}{2\varepsilon}}$ of P_ε , normalized at x_A , is $e^{-\frac{1}{\varepsilon} d_{\text{Ag}}(x_A, \cdot)}$ on the basin of x_A , and the Arrhenius factor $e^{-\frac{L(z^*) - L(x_A)}{\varepsilon}}$ is the tunneling factor $e^{-\frac{2}{\varepsilon} d_{\text{Ag}}(x_A, z^*)}$.

ⁱH. Eyring, *The activated complex in chemical reactions*. J. Chem. Phys. 3 (1935), 107–115.

ⁱⁱH. Kramers, *Brownian motion in a field of force and the diffusion model of chemical reactions*. Physica 7 (1940), 284–304.

8.5. Capacities

The approach of [BEGK] expresses mean hitting times through capacities, which are estimated by variational principles. The Dirichlet form of the dynamics is

$$\mathcal{E}_\varepsilon(f) = \varepsilon \int_{\mathbb{T}^d} |\nabla f|^2 d\pi_\varepsilon \quad \text{for } f \in H^1(\mathbb{T}^d).$$

Definition (Capacity). Let $A, B \subset \mathbb{T}^d$ be disjoint closed sets with non-empty interiors. The capacity of the pair (A, B) is

$$\text{cap}_\varepsilon(A, B) = \inf \left\{ \mathcal{E}_\varepsilon(f) : f \in H^1(\mathbb{T}^d), f = 1 \text{ a.e. on } A, f = 0 \text{ a.e. on } B \right\}.$$

In electrostatics, the capacity of a capacitor formed by two conductors A and B is the energy $\int |\nabla f|^2$ of the electrostatic potential f which equals 1 on A and 0 on B . This potential minimizes the energy. Here the medium has the inhomogeneous conductivity $\varepsilon e^{-\frac{L}{\varepsilon}}$. So $\text{cap}_\varepsilon(A, B)$ is small if every path from A to B passes through a region where L is large. The infimum is attained at the equilibrium potential $q_{A,B}$, and $0 \leq q_{A,B} \leq 1$. This follows by truncation and the strict convexity of $\mathcal{E}_\varepsilon$; see [BEGK, Section 2].

Theorem (Bovier-Eckhoff-Gaynard-Klein). *Let $A, B \subset \mathbb{T}^d$ be disjoint closed sets whose complements have C^1 boundaries. Then there is a probability measure $\nu_{A,B}$ on ∂A , the normalized equilibrium measure, such that*

$$\int_{\partial A} \mathbb{E}_w[\tau_B] \nu_{A,B}(dw) = \frac{1}{\text{cap}_\varepsilon(A, B)} \int_{\mathbb{T}^d \setminus B} q_{A,B} d\pi_\varepsilon.$$

This is [BEGK, (2.27)], stated there in a domain of $\mathbb{R}^d$ with a small ball in place of A . The formula explains the heuristic for mean transition times. The function $q_{A,B}$ is close to 1 on the well of A and close to 0 on the well of B . So the numerator is essentially the Gibbs mass of the starting well, while the denominator is the conductance of the narrowest passage between the wells.

Remark (The formula as Green's identity). The formula of Bovier-Eckhoff-Gaynard-Klein is Green's identity for the weighted Laplacian $\mathcal{L}_\varepsilon$. Suppose, for simplicity, that ∂A and ∂B are smooth, and write

$$\Omega = \mathbb{T}^d \setminus (A \cup B) \quad \text{and} \quad q = q_{A,B}.$$

By the proof of Proposition 8.1, $e^{-\frac{L}{\varepsilon}} \mathcal{L}_\varepsilon g = \varepsilon \text{div}(e^{-\frac{L}{\varepsilon}} \nabla g)$. Let $\mathbf{n}$ be the unit normal on ∂A pointing into A , and let dS be the surface measure on ∂A . Since $q \leq 1$ in Ω and $q = 1$ on ∂A , we have $\partial_{\mathbf{n}} q \geq 0$ on ∂A . The equilibrium measure is the flux of the current $\varepsilon e^{-\frac{L}{\varepsilon}} \nabla q$ into A , that is,

$$e_{A,B}(dS) = \varepsilon \partial_{\mathbf{n}} q e^{-\frac{L}{\varepsilon}} dS.$$

First, $\text{div}(e^{-\frac{L}{\varepsilon}} q \nabla q) = e^{-\frac{L}{\varepsilon}} |\nabla q|^2$ in Ω , $\nabla q = 0$ in the interiors of A and B , and $q = 0$ on ∂B . So the divergence theorem in Ω gives

$$\text{cap}_\varepsilon(A, B) = \mathcal{E}_\varepsilon(q) = \varepsilon \int_{\partial A} \partial_{\mathbf{n}} q e^{-\frac{L}{\varepsilon}} dS = e_{A,B}(\partial A).$$

Second, $\mathcal{L}_\varepsilon q = 0$ and $\mathcal{L}_\varepsilon w_B = -1$ in Ω , and $q = w_B = 0$ on ∂B . So Green's identity in Ω gives

$$-\int_{\Omega} q d\pi_\varepsilon = \int_{\Omega} (q \mathcal{L}_\varepsilon w_B - w_B \mathcal{L}_\varepsilon q) d\pi_\varepsilon = \varepsilon \int_{\partial A} (\partial_{\mathbf{n}} w_B - w_B \partial_{\mathbf{n}} q) e^{-\frac{L}{\varepsilon}} dS.$$

Since $\mathcal{L}_\varepsilon w_B = -1$ also in A , the divergence theorem in A gives

$$\varepsilon \int_{\partial A} \partial_{\mathbf{n}} w_B e^{-\frac{L}{\varepsilon}} dS = - \int_A \mathcal{L}_\varepsilon w_B d\pi_\varepsilon = \pi_\varepsilon(A).$$

Since $q = 1$ on A , we obtain

$$\int_{\partial A} w_B de_{A,B} = \pi_\varepsilon(A) + \int_\Omega q d\pi_\varepsilon = \int_{\mathbb{T}^d \setminus B} q d\pi_\varepsilon.$$

Dividing by $\text{cap}_\varepsilon(A, B) = e_{A,B}(\partial A)$ gives the formula, with

$$\nu_{A,B} = \frac{e_{A,B}}{\text{cap}_\varepsilon(A, B)}.$$

So the proof uses only the boundary value problems for $q_{A,B}$ and w_B .

We explain the Eyring-Kramers formula from this, since the same computation underlies Chapter 9. By Laplace's method, as in Proposition 6.1, the Gibbs mass of a small ball A around a non-degenerate minimum x_A is

$$\pi_\varepsilon(A) = (2\pi\varepsilon)^{\frac{d}{2}} \left(\det \nabla^2 L(x_A) \right)^{-\frac{1}{2}} e^{-\frac{L(x_A)}{\varepsilon}} (1 + o(1)).$$

For the capacity, let $h = L(z^*)$, and let e_1 be the unstable direction at z^* . Consider first the one-dimensional problem. If f is a function on $[-\delta, \delta]$ with $f(-\delta) = 1$ and $f(\delta) = 0$, then, by the Cauchy-Schwarz inequality,

$$1 \leq \left(\int_{-\delta}^{\delta} |f'(s)| ds \right)^2 \leq \int_{-\delta}^{\delta} |f'(s)|^2 e^{-\frac{L(z^* + se_1)}{\varepsilon}} ds \int_{-\delta}^{\delta} e^{\frac{L(z^* + se_1)}{\varepsilon}} ds,$$

with equality if f' is proportional to $e^{\frac{L(z^* + se_1)}{\varepsilon}}$. Since

$$L(z^* + se_1) = h - \frac{\mu_-}{2} s^2 + O(|s|^3),$$

Laplace's method shows that the one-dimensional capacity through the saddle point is

$$\varepsilon e^{-\frac{h}{\varepsilon}} \sqrt{\frac{\mu_-}{2\pi\varepsilon}} (1 + o(1)).$$

In the $d - 1$ stable directions, L increases quadratically. Integrating the one-dimensional capacity against the transversal Gaussian weight produces the factor $(2\pi\varepsilon)^{\frac{d-1}{2}} (\det H')^{-\frac{1}{2}}$, where H' is the restriction of $\nabla^2 L(z^*)$ to $e_1^\perp$. The product is

$$\kappa_\varepsilon = \frac{\mu_-}{2\pi} (2\pi\varepsilon)^{\frac{d}{2}} \left| \det \nabla^2 L(z^*) \right|^{-\frac{1}{2}} e^{-\frac{h}{\varepsilon}}, \quad (8.2)$$

and the ratio $\frac{\pi_\varepsilon(A)}{\kappa_\varepsilon}$ is the Eyring-Kramers formula.

8.6. The semiclassical point of view

The generator $\mathcal{L}_\varepsilon$ is unitarily equivalent to a semiclassical Schrödinger operator, with the temperature ε in the role of the semiclassical parameter.

Lemma 8.4. *The map $f \mapsto e^{-\frac{L}{2\varepsilon}} f$ is a unitary operator from $L^2(\mathbb{T}^d, \pi_\varepsilon)$ onto $L^2(\mathbb{T}^d, dw)$. For $v \in C^2(\mathbb{T}^d)$,*

$$e^{-\frac{L}{2\varepsilon}} (-\varepsilon \mathcal{L}_\varepsilon) (e^{\frac{L}{2\varepsilon}} v) = P_\varepsilon v, \quad \text{where } P_\varepsilon = -\varepsilon^2 \Delta + \frac{1}{4} |\nabla L|^2 - \frac{\varepsilon}{2} \Delta L.$$

PROOF. The first claim is clear from $\pi_\varepsilon(dw) = e^{-\frac{L}{\varepsilon}}dw$. Let $u = e^{\frac{L}{2\varepsilon}}v$. Then

$$\nabla u = e^{\frac{L}{2\varepsilon}} \left(\nabla v + \frac{v \nabla L}{2\varepsilon} \right)$$

and

$$\Delta u = e^{\frac{L}{2\varepsilon}} \left(\Delta v + \frac{\nabla L \cdot \nabla v}{\varepsilon} + \frac{v \Delta L}{2\varepsilon} + \frac{v |\nabla L|^2}{4\varepsilon^2} \right).$$

Hence

$$-\varepsilon \mathcal{L}_\varepsilon u = -\varepsilon^2 \Delta u + \varepsilon \nabla L \cdot \nabla u = e^{\frac{L}{2\varepsilon}} \left(-\varepsilon^2 \Delta v + \frac{1}{4} |\nabla L|^2 v - \frac{\varepsilon}{2} v \Delta L \right).$$

□

The operator P_ε is the Witten Laplacian on functions associated with $\frac{L}{2}$, introduced by Wittenⁱ in his analytic approach to Morse theory. See also the book of Helffer-Nierⁱⁱ and Delande [D, Section 1.1]. By Proposition 8.1, P_ε is non-negative, and $P_\varepsilon e^{-\frac{L}{2\varepsilon}} = 0$, i.e., $e^{-\frac{L}{2\varepsilon}}$ is its ground state. Since $\mathbb{T}^d$ is connected, $\mathcal{E}_\varepsilon(f) = 0$ only for constant f , so the ground state is simple. Since $-\mathcal{L}_\varepsilon$ is unitarily equivalent to $\frac{1}{\varepsilon} P_\varepsilon$, the spectral gap of the generator, i.e., the smallest non-zero eigenvalue of $-\mathcal{L}_\varepsilon$, is $\frac{1}{\varepsilon}$ times the smallest non-zero eigenvalue of P_ε . To leading order, the potential of P_ε is $\frac{1}{4} |\nabla L|^2$, which vanishes exactly at the critical points of L . So the low-lying spectrum of P_ε is decided at the critical points.

Remark (The harmonic approximation). Let z be a non-degenerate critical point of L , and let $\mu_1, \dots, \mu_d$ be the eigenvalues of $\nabla^2 L(z)$. If we replace L by its Taylor polynomial of order two at z , then P_ε becomes the harmonic oscillator

$$-\varepsilon^2 \Delta + \frac{1}{4} \left| \nabla^2 L(z)(w - z) \right|^2 - \frac{\varepsilon}{2} \sum_{j=1}^d \mu_j,$$

whose eigenvalues are

$$\varepsilon \sum_{j=1}^d \left(|\mu_j| n_j + \frac{|\mu_j| - \mu_j}{2} \right) \quad \text{for } n_1, \dots, n_d \in \{0, 1, 2, \dots\}.$$

So its lowest eigenvalue is 0 if z is a local minimum, and $\varepsilon \sum_{\mu_j < 0} |\mu_j| > 0$ otherwise.

For a Morse function L , the harmonic approximation of Simonⁱⁱⁱ and of Helffer-Sjöstrand^{iv} shows that, for each k , the k -th eigenvalue of $\frac{1}{\varepsilon} P_\varepsilon$ converges as $\varepsilon \rightarrow 0$ to the k -th smallest of the numbers $\sum_j (|\mu_j| n_j + \frac{|\mu_j| - \mu_j}{2})$ over all critical points z and all $(n_1, \dots, n_d)$. In particular, if $c > 0$ is smaller than the smallest positive one of these numbers, then, for small ε , P_ε has as many eigenvalues in $[0, c\varepsilon]$ as L has local minima. These eigenvalues are in fact exponentially small. Under generic assumptions, Helffer-Klein-Nier^v and Bovier-Gaynard-Klein^{vi} proved that

ⁱE. Witten, *Supersymmetry and Morse theory*. J. Differential Geom. 17 (1982), 661–692.

ⁱⁱB. Helffer and F. Nier, *Hypoelliptic Estimates and Spectral Theory for Fokker-Planck Operators and Witten Laplacians*. Springer, 2005.

ⁱⁱⁱB. Simon, *Semiclassical analysis of low lying eigenvalues. I. Non-degenerate minima: asymptotic expansions*. Ann. Inst. H. Poincaré Sect. A (N.S.) 38 (1983), 295–308.

^{iv}B. Helffer and J. Sjöstrand, *Multiple wells in the semi-classical limit. I*. Comm. Partial Differential Equations 9 (1984), 337–408.

^vB. Helffer, M. Klein, and F. Nier, *Quantitative analysis of metastability in reversible diffusion processes via a Witten complex approach*. Mat. Contemp. 26 (2004), 41–85.

^{vi}A. Bovier, V. Gaynard, and M. Klein, *Metastability in reversible diffusion processes. II. Precise asymptotics for small eigenvalues*. J. Eur. Math. Soc. (JEMS) 7 (2005), 69–99.

the non-zero ones are ε times the inverses of the Eyring-Kramers mean times of the transitions from the local minima to the deeper ones, up to factors $1 + o(1)$. For a singular well, $\nabla^2 L$ degenerates along the well, and the potential $\frac{1}{4}|\nabla L|^2$ vanishes on the whole well, which may be positive-dimensional. So the harmonic approximation is not available, and the low-lying spectrum of P_ε reflects the singular geometry of the well. For instance, the size of the spectral gap in the example of Chapter 10 is Problem 10.1(b).

Remark (Spectral gap and relaxation). Let $U \subset \mathbb{T}^d$ be a connected open set with smooth boundary, and let

$$\hat{\pi} = \frac{\mathbb{1}_U \pi_\varepsilon}{\pi_\varepsilon(U)}$$

be the normalized Gibbs measure on U . The quadratic form $\varepsilon \int_U |\nabla f|^2 d\hat{\pi}$ on $H^1(U)$, with no boundary condition, defines a non-negative self-adjoint operator $-\mathcal{L}_\varepsilon^N$ on $L^2(U, \hat{\pi})$. It is $-\mathcal{L}_\varepsilon$ with the Neumann condition $\partial_{\mathbf{n}} f = 0$ on ∂U , where $\mathbf{n}$ is the unit normal. It generates the dynamics reflected at ∂U , i.e., the Langevin dynamics pushed back into U whenever it reaches ∂U . Its kernel consists of the constants. By the min-max principle, its spectral gap γ_ε , i.e., its smallest non-zero eigenvalue, is

$$\gamma_\varepsilon = \inf \left\{ \frac{\varepsilon \int_U |\nabla f|^2 d\hat{\pi}}{\int_U |f - \hat{\pi}(f)|^2 d\hat{\pi}} : f \in H^1(U) \text{ is not constant} \right\} \quad \text{with } \hat{\pi}(f) = \int_U f d\hat{\pi}.$$

So γ_ε is the best constant in the Poincaré inequality for $\hat{\pi}$. Since $e^{t\mathcal{L}_\varepsilon^N} 1 = 1$, the spectral theorem gives

$$\left\| e^{t\mathcal{L}_\varepsilon^N} f - \hat{\pi}(f) \right\|_{L^2(U, \hat{\pi})} \leq e^{-\gamma_\varepsilon t} \|f - \hat{\pi}(f)\|_{L^2(U, \hat{\pi})}.$$

So, after a time large compared with γ_ε^{-1} , the expectations $\mathbb{E}_w[f(w_t)]$ for the reflected dynamics no longer depend on the starting point w , in the L^2 sense. This is why γ_ε^{-1} is called the relaxation time, and why a lower bound on γ_ε , polynomial in ε , expresses the equilibration within a well. Upper bounds are easier: for every non-constant $f \in H^1(U)$, the quotient in the infimum above is an upper bound for γ_ε . For $U = \mathbb{T}^d$, there is no boundary, and γ_ε is the spectral gap of the generator above. In this case, suppose that the law of w_0 has the density $g_0 e^{-\frac{L}{\varepsilon}}$ with $g_0 \in L^2(\mathbb{T}^d, \pi_\varepsilon)$. By the Fokker-Planck equation and the identity in the proof of Proposition 8.1, the law of w_t has the density $g_t e^{-\frac{L}{\varepsilon}}$ with

$$g_t = e^{t\mathcal{L}_\varepsilon} g_0.$$

Since $\int g_0 d\pi_\varepsilon = 1$, we have $\hat{\pi}(g_0) = \pi_\varepsilon(\mathbb{T}^d)^{-1}$, and the bound above gives

$$\left\| g_t - \pi_\varepsilon(\mathbb{T}^d)^{-1} \right\|_{L^2(\pi_\varepsilon)} \leq e^{-\gamma_\varepsilon t} \left\| g_0 - \pi_\varepsilon(\mathbb{T}^d)^{-1} \right\|_{L^2(\pi_\varepsilon)}.$$

So the law of w_t approaches the normalized Gibbs measure on the time scale γ_ε^{-1} .

CHAPTER 9

Eyring-Kramers laws for singular wells

In Chapter 8, the wells of the potential were non-degenerate minima. For the loss of a neural network, the wells are positive-dimensional singular sets of optimal parameters (Chapter 2). How does a singular well change the Eyring-Kramers formula? In this chapter, we state the problems MAIS-O73 and MAIS-O74, prove the conjectured law for the mean transition time started from the equilibrium measure, compute the constant for monomial wells, and explain what remains open. Our presentation mainly follows the author's work [H].

9.1. The problems

Let $L : \mathbb{T}^d \rightarrow [0, \infty)$ be real-analytic. Suppose that its zero set consists of two disjoint compact connected components, the wells, i.e.,

$$W_0 = L^{-1}(0) = K_A \sqcup K_B.$$

To each well K , we attach the pair (λ_K, m_K) : the minimal local learning coefficient over K , and then the maximal multiplicity among the points of K attaining it. Let h be the communication height between K_A and K_B (Chapter 8). For a small $\rho > 0$, let τ_B be the hitting time of

$$\{w : \text{dist}(w, K_B) \leq \rho\}.$$

Let $\mathcal{B}_A$ be the basin of K_A , i.e., the set of points from which gradient flow converges to a point of K_A .

Problem 9.1 (MAIS-O73). *Determine the asymptotics of $\mathbb{E}_w[\tau_B]$ as $\varepsilon \rightarrow 0$, uniformly for w in compact subsets of the interior of $\mathcal{B}_A \cap \{L < h\}$, in terms of the local data of L at K_A and the geometry of L at the gate between the wells.*

On the exponential scale, the answer $e^{\frac{h}{\varepsilon}}$ is supplied by the Freidlin-Wentzell theory (Chapter 8). So the problem is the prefactor. The general problem allows degenerate gates. The problem MAIS-O74 is the case of a single Morse saddle.

Conjecture (MAIS-O74). *In the setting above, suppose that the infimum h is attained at a single non-degenerate critical point z^* of index one. Then*

$$\mathbb{E}_w[\tau_B] = C\varepsilon^{\lambda_A - \frac{d}{2}} \left(\log \frac{1}{\varepsilon} \right)^{m_A - 1} e^{\frac{h}{\varepsilon}} (1 + o(1)),$$

uniformly for w in compact subsets of the interior of $\mathcal{B}_A \cap \{L < h\}$, where $C > 0$ is explicit in the leading coefficient of the volume asymptotics of L at K_A and the Hessian data of L at z^ .*

The heuristic is the ratio in the formula of Bovier-Eckhoff-Gayrard-Klein (Section 8.5): the mean transition time is the Gibbs mass of the starting well divided by the capacity of the gate. The mass is of order $\varepsilon^{\lambda_A} (\log \frac{1}{\varepsilon})^{m_A - 1}$ by the volume law, and the capacity of a Morse gate is of order $\varepsilon^{\frac{d}{2}} e^{-\frac{h}{\varepsilon}}$, as in (8.2). When K_A is a non-degenerate minimum, $\lambda_A = \frac{d}{2}$ and $m_A = 1$, and the conjecture is the Eyring-Kramers formula. In dimension one, the conjecture holds for each starting point by Theorem 8.3. In higher dimensions, the existing theorems assume non-degenerate wells

(Bovier-Eckhoff-Gaynard-Klein [BEGK]), isolated degenerate minima (Avelin-Julin-Viitasaariⁱ), Morse-Bott wells (Assal-Bony-Michelⁱⁱ), or isolated critical points with diagonal monomial normal forms (Delande [D]). Here a Morse-Bott well is a smooth submanifold along which the Hessian is non-degenerate in the normal directions. None of these results covers a positive-dimensional singular well.

We formulate the hypothesis on the gate as follows. Let G_A and G_B be the connected components of $\{L < h\}$ containing K_A and K_B , respectively. We assume that z^* is a Morse saddle with $L(z^*) = h$, and we write $-\mu_-$ for the negative eigenvalue of $\nabla^2 L(z^*)$. We assume, moreover, that the two sides of z^* in the unstable direction lie in G_A and G_B , respectively, and that every path from K_A to K_B along which L stays slightly above h passes near z^* . See [H, Section 1.1] for the precise statements.

9.2. The Gibbs mass of a singular well

The volume law of Chapter 6 transfers to the Gibbs mass by the computation in Proposition 6.2, with $\frac{1}{\varepsilon}$ in place of the sample size n .

Proposition 9.1. *Suppose that, for a closed neighborhood A of K_A disjoint from K_B ,*

$$V_A(\varepsilon) = \text{Leb}(\{w \in A : L(w) < \varepsilon\}) = c_A \varepsilon^{\lambda_A} \left(\log \frac{1}{\varepsilon} \right)^{m_A-1} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0^+.$$

Then

$$\pi_\varepsilon(A) = \Gamma(\lambda_A + 1) c_A \varepsilon^{\lambda_A} \left(\log \frac{1}{\varepsilon} \right)^{m_A-1} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0.$$

PROOF. As in Proposition 6.2,

$$\pi_\varepsilon(A) = \int_0^\infty e^{-u} V_A(\varepsilon u) du.$$

The same dominated convergence argument applies, with the bound $V_A \leq (2\pi)^d$ in place of $V \leq 1$. $\square$

So the Gibbs mass is governed by the same volume law as the Bayesian free energy, and selection at equilibrium is Watanabe's computation. For real-analytic L , the volume law holds with (λ_A, m_A) as above. Indeed, let $\psi \geq 0$ be a smooth cutoff with $\text{supp } \psi \subset A$ and $\psi = 1$ near K_A . Since L is bounded below by a positive constant on $A \setminus \{\psi = 1\}^\circ$, we have

$$V_A(\varepsilon) = \int_{\{L < \varepsilon\}} \psi dw \quad \text{for small } \varepsilon.$$

So the volume law follows by resolution of singularities, as in Section 6.3, together with the localization of the learning coefficient to the points of K_A (Watanabe [W]). In [H], the volume law is taken as a hypothesis. By Lemma 6.4, $\lambda_A \leq \frac{d}{2}$.

ⁱB. Avelin, V. Julin, and L. Viitasaari, *Geometric characterization of the Eyring-Kramers formula*. Comm. Math. Phys. 404 (2023), 401–437.

ⁱⁱM. Assal, J.-F. Bony, and L. Michel, *Metastable diffusions with degenerate drifts*. Ann. Inst. Fourier (Grenoble) 75 (2025), 1–33.

9.3. Capacity through a Morse gate

We call two closed sets $A, B \subset \mathbb{T}^d$ an admissible pair if they are connected, K_A and K_B lie in their interiors, respectively, and

$$\max_{A \cup B} L < h.$$

For instance, the closed ρ -neighborhoods of K_A and K_B form an admissible pair for small ρ .

Theorem 9.2 ([H]). *Under the hypothesis on the gate, for every admissible pair (A, B) ,*

$$\text{cap}_\varepsilon(A, B) = \kappa_\varepsilon (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0,$$

where κ_ε is given by (8.2).

The theorem requires nothing of the wells: the capacity is decided at the gate. We outline the proof. For the upper bound, one constructs a test function f in the definition of the capacity. Let e_1 be a unit eigenvector of $\nabla^2 L(z^*)$ for the eigenvalue $-\mu_-$. Near z^* , in coordinates $w = z^* + y_1 e_1 + y'$ with $y' \perp e_1$, let f depend only on y_1 , through the optimal one-dimensional profile

$$g(y_1) = \frac{\int_{y_1}^\delta e^{-\frac{\mu_- s^2}{2\varepsilon}} ds}{\int_{-\delta}^\delta e^{-\frac{\mu_- s^2}{2\varepsilon}} ds}.$$

Away from z^* , let $f = 1$ on the side of K_A and $f = 0$ on the side of K_B . The hypothesis on the gate guarantees that the two sides can be separated by a region where $L > h$, and there the energy of f is negligible. The energy of f near z^* is the Gaussian computation in Section 8.5.

For the lower bound, the energy of every admissible function is bounded from below along a family of curves, by the following lemma, whose idea goes back to [BEGK, Proposition 4.7].

Lemma 9.3 (Tube lower bound). *Let $\mathcal{W} \subset \mathbb{R}^{d-1}$ be a bounded open set, and let $\Psi : [0, S] \times \mathcal{W} \rightarrow \mathbb{T}^d$ be an injective C^1 map whose Jacobian determinant J is positive, such that $\Psi(0, v) \in A$ and $\Psi(S, v) \in B$ for $v \in \mathcal{W}$. Then, for every $f \in C^1(\mathbb{T}^d)$ with $f = 1$ on A and $f = 0$ on B ,*

$$\mathcal{E}_\varepsilon(f) \geq \varepsilon \int_{\mathcal{W}} \frac{dv}{R(v)}, \quad \text{where } R(v) = \int_0^S \frac{|\partial_s \Psi(s, v)|^2}{J(s, v)} e^{\frac{L(\Psi(s, v))}{\varepsilon}} ds.$$

PROOF. By the change of variables $w = \Psi(s, v)$,

$$\mathcal{E}_\varepsilon(f) \geq \varepsilon \int_{\mathcal{W}} \int_0^S |\nabla f(\Psi)|^2 e^{-\frac{L(\Psi)}{\varepsilon}} J ds dv.$$

For fixed v , let $F(s) = f(\Psi(s, v))$. Then $|F'(s)| \leq |\nabla f(\Psi)| |\partial_s \Psi|$. By the Cauchy-Schwarz inequality,

$$\begin{aligned} 1 = F(0) - F(S) &\leq \int_0^S |F'| ds \leq \left(\int_0^S \frac{|F'|^2 J}{|\partial_s \Psi|^2} e^{-\frac{L(\Psi)}{\varepsilon}} ds \right)^{\frac{1}{2}} R(v)^{\frac{1}{2}} \\ &\leq \left(\int_0^S |\nabla f(\Psi)|^2 e^{-\frac{L(\Psi)}{\varepsilon}} J ds \right)^{\frac{1}{2}} R(v)^{\frac{1}{2}}. \end{aligned}$$

Squaring, dividing by $R(v)$, and integrating over v completes the proof. $\square$

To obtain the sharp lower bound, one takes a tube of straight fibers $\Psi(s, v) = z^* + s e_1 + v$ through the saddle point, with $v \perp e_1$ ranging over a ball of radius $\varepsilon^{\frac{1}{2}} \log \frac{1}{\varepsilon}$, and continues the

fibers through the wells to A and B ; see Figure 9.1. Near z^* , we have $J = 1$ and $|\partial_s \Psi| = 1$. Let H' be the restriction of $\nabla^2 L(z^*)$ to $e_1^\perp$, as in Section 8.5. Then

$$R(v) \approx e^{\frac{h}{\varepsilon}} e^{\frac{v \cdot H' v}{2\varepsilon}} \sqrt{\frac{2\pi\varepsilon}{\mu_-}}.$$

So the integral $\varepsilon \int \frac{dv}{R(v)}$ is the Gaussian computation of κ_ε in Section 8.5. The continuation through the wells costs exponentially less, because $L < h$ there. In **[H]**, the continuation is made of translates of one polygonal path, and weights compensate for its self-intersections. No information on the wells is needed.

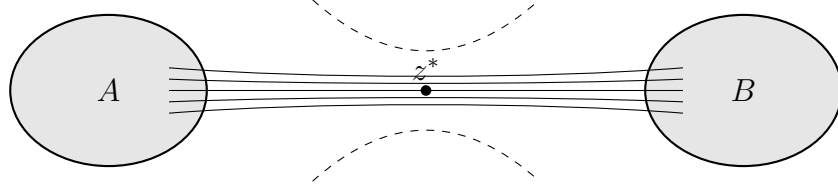


FIGURE 9.1. A tube of fibers from A to B through the saddle point z^* ; the dashed curves are level curves of L at a level slightly above h

9.4. An averaged singular Eyring-Kramers law

Combining Proposition 9.1, Theorem 9.2, and the formula of Bovier-Eckhoff-Gayard-Klein, we obtain the conjectured law for the mean transition time started from the normalized equilibrium measure $\nu_{A,B}$ of Section 8.5.

Theorem 9.4 ([H]). *Suppose that K_A satisfies the volume law in Proposition 9.1, and that the gate is a Morse saddle as above. Let (A, B) be an admissible pair whose complements have C^1 boundaries. Then*

$$\int_{\partial A} \mathbb{E}_w [\tau_B] \nu_{A,B}(dw) = C_* \varepsilon^{\lambda_A - \frac{d}{2}} \left(\log \frac{1}{\varepsilon} \right)^{m_A - 1} e^{\frac{h}{\varepsilon}} (1 + o(1)) \quad \text{as } \varepsilon \rightarrow 0,$$

where

$$C_* = (2\pi)^{1 - \frac{d}{2}} \frac{\Gamma(\lambda_A + 1) c_A |\det \nabla^2 L(z^*)|^{\frac{1}{2}}}{\mu_-}.$$

PROOF. By the formula of Bovier-Eckhoff-Gayard-Klein, the left-hand side equals

$$\frac{1}{\text{cap}_\varepsilon(A, B)} \int_{\mathbb{T}^d \setminus B} q_{A,B} d\pi_\varepsilon.$$

Since $q_{A,B} = 1$ on A and $0 \leq q_{A,B} \leq 1$,

$$0 \leq \int_{\mathbb{T}^d \setminus B} q_{A,B} d\pi_\varepsilon - \pi_\varepsilon(A) = \int_{\mathbb{T}^d \setminus (A \cup B)} q_{A,B} d\pi_\varepsilon \leq (2\pi)^d e^{-\frac{\ell}{\varepsilon}},$$

where $\ell > 0$ is the minimum of L on the compact set $\mathbb{T}^d \setminus (A^\circ \cup B^\circ)$, which avoids W_0 . This error is negligible compared to $\pi_\varepsilon(A)$, which is given by Proposition 9.1. Dividing by $\text{cap}_\varepsilon(A, B) = \kappa_\varepsilon(1 + o(1))$ from Theorem 9.2 gives the theorem. $\square$

Remark. When $K_A = \{x_A\}$ is a non-degenerate minimum, the regular case of the examples in Section 6.3 gives

$$\Gamma\left(\frac{d}{2} + 1\right) c_A = (2\pi)^{\frac{d}{2}} \left(\det \nabla^2 L(x_A) \right)^{-\frac{1}{2}}.$$

So C_* is the Eyring-Kramers prefactor. Since $\lambda_A \leq \frac{d}{2}$, the factor $\varepsilon^{\lambda_A - \frac{d}{2}}$ is never small. So, compared with a non-degenerate well, a well with $\lambda_A < \frac{d}{2}$ lengthens the transition time by this power of ε . Moreover, the capacity is symmetric in (A, B) . So if K_B also satisfies a volume law, then the mean transition time from K_A to K_B divided by the one from K_B to K_A is

$$\frac{\Gamma(\lambda_A + 1) c_A}{\Gamma(\lambda_B + 1) c_B} \varepsilon^{\lambda_A - \lambda_B} \left(\log \frac{1}{\varepsilon} \right)^{m_A - m_B} (1 + o(1)),$$

i.e., the ratio of the Gibbs masses of the two wells, as one expects from reversibility.

The constant c_A is explicit for monomial wells, a natural first target. Let $k_1, \dots, k_d \geq 1$ be integers, and set

$$F(w) = \prod_{j=1}^d \sin^{2k_j} \left(\frac{w_j}{2} \right), \quad \varkappa = \max_j k_j, \quad I = \{j : k_j = \varkappa\}, \quad \text{and} \quad \lambda = \frac{1}{2\varkappa}.$$

The zero set of F is the union of the coordinate subtori $\{w_j = 0\}$, which is singular along its crossings. Near the origin, F is $\prod_j w_j^{2k_j}$ times a positive analytic factor. Write $w = (w_I, w')$ with $w' = (w_j)_{j \notin I}$, and let $m_0 = \text{Card } I$. Here m_0 plays the role of the multiplicity.

Proposition 9.5 ([H]). *Suppose that $K_A = F^{-1}(0)$ and $L = FG$ on a neighborhood of K_A , where G is continuous and positive there. Then the volume law holds with $\lambda_A = \lambda$, $m_A = m_0$, and*

$$c_A = \frac{4^{m_0} \lambda^{m_0-1}}{(m_0 - 1)!} \int_{\mathbb{T}^{d-m_0}} G(0_I, w')^{-\lambda} \prod_{j \notin I} \left| \sin \frac{w_j}{2} \right|^{-\frac{k_j}{\varkappa}} dw'.$$

The proof in [H] is probabilistic. If w is uniformly distributed on $\mathbb{T}^d$, then the variables $U_j = -\log \left| \sin \frac{w_j}{2} \right|$ are independent, with the tails

$$\mathbb{P}[U_j > s] = \frac{2}{\pi} \arcsin(e^{-s}),$$

and $\{F < \varepsilon\}$ is the event

$$\sum_{j=1}^d 2k_j U_j > \log \frac{1}{\varepsilon}.$$

So the volume law reduces to the tail of a sum of independent random variables. Since $\mathbb{P}[2k_j U_j > s]$ is of order $e^{-\frac{s}{2k_j}}$, the terms with $j \in I$ dominate. For $j \in I$, the tail of $2\varkappa U_j$ is comparable to that of an exponential random variable with mean $\frac{1}{\lambda}$. So, as for a sum of m_0 independent such variables, the sum of these m_0 terms has a tail of order $s^{m_0-1} e^{-\lambda s}$. With $s = \log \frac{1}{\varepsilon}$, this is the order $\varepsilon^\lambda (\log \frac{1}{\varepsilon})^{m_0-1}$ of the volume law.

Example. Let $d = 2$ and $G \equiv 1$.

- (1) If $(k_1, k_2) = (1, 1)$, then $\lambda_A = \frac{1}{2}$, $m_A = 2$, and $c_A = 8$. So the transition time carries the factor $\varepsilon^{-\frac{1}{2}} \log \frac{1}{\varepsilon}$.
- (2) If $(k_1, k_2) = (2, 1)$, then $\lambda_A = \frac{1}{4}$, $m_A = 1$, and

$$c_A = \frac{8\sqrt{\pi} \Gamma\left(\frac{1}{4}\right)}{\Gamma\left(\frac{3}{4}\right)}.$$

So the transition time carries the factor $\varepsilon^{-\frac{3}{4}}$.

Remark (The prefactor is not local). Let $(k_1, k_2) = (2, 1)$ as in Example (2), but let G be general. Then $I = \{1\}$, and

$$c_A = 4 \int_{-\pi}^{\pi} G(0, w_2)^{-\frac{1}{4}} \left| \sin \frac{w_2}{2} \right|^{-\frac{1}{2}} dw_2.$$

So c_A depends on the values of G along the whole circle $\{w_1 = 0\}$. Hence the prefactor C_* is not determined by the germ of L at any single point of K_A . It is a global quantity along the deepest stratum of the well.

Remark (Where the Gibbs mass sits). By Lemma 8.4, $\pi_\varepsilon(dw) = \left(e^{-\frac{L}{2\varepsilon}}\right)^2 dw$ is the square of the ground state of the Witten Laplacian P_ε . So the weak limits of the normalized measures $\pi_\varepsilon(A)^{-1} \mathbf{1}_A \pi_\varepsilon$ as $\varepsilon \rightarrow 0$ play the role of the semiclassical measures of a sequence of eigenfunctions. The computation behind Proposition 9.5, with a continuous function f inserted in the integrand, identifies the limit on a monomial well. It is the probability measure on the subtorus $\{w_I = 0\}$ whose density with respect to dw' is proportional to

$$G(0_I, w')^{-\lambda} \prod_{j \notin I} \left| \sin \frac{w_j}{2} \right|^{-\frac{k_j}{\varkappa}}.$$

So the Gibbs mass concentrates on the deepest stratum of the well, where the local learning coefficient is minimal and the multiplicity is maximal. It is not uniform there: it accumulates near the crossings with the other coordinate subtori.

9.5. The pointwise problem

Theorem 9.4 concerns the mean transition time when the starting point is distributed according to $\nu_{A,B}$. The conjecture asks for each starting point w . We explain what is missing [H, Section 9]. Let

$$T_\varepsilon = \int_{\partial A} \mathbb{E}_y[\tau_B] \nu_{A,B}(dy).$$

For $w \notin A \cup B$, the strong Markov property at the hitting time of $A \cup B$ gives

$$\mathbb{E}_w[\tau_B] = \mathbb{E}_w[\tau_{A \cup B}] + \mathbb{E}_w[\mathbf{1}_{\{\tau_A < \tau_B\}} \mathbb{E}_{w_{\tau_A}}[\tau_B]].$$

Since $w_{\tau_A} \in \partial A$ and

$$\inf_{\partial A} \mathbb{E}[\tau_B] \leq T_\varepsilon \leq \sup_{\partial A} \mathbb{E}[\tau_B],$$

the pointwise law on a compact set $D \subset \mathbb{T}^d \setminus (A \cup B)$ follows from Theorem 9.4 and the three estimates

- (a) $\sup_{w \in D} \mathbb{E}_w[\tau_{A \cup B}] = o(T_\varepsilon)$,
- (b) $\sup_{w \in D} \mathbb{P}_w[\tau_B < \tau_A] = o(1)$,
- (c) $\sup_{y, y' \in \partial A} |\mathbb{E}_y[\tau_B] - \mathbb{E}_{y'}[\tau_B]| = o(T_\varepsilon)$.

Recall from Chapter 8 that $w_B(y) = \mathbb{E}_y[\tau_B]$ solves the Dirichlet problem

$$-\mathcal{L}_\varepsilon w_B = 1 \text{ in } \mathbb{T}^d \setminus B, \quad w_B = 0 \text{ on } B.$$

Remark (The strong Markov property as superposition). Let $\Omega = \mathbb{T}^d \setminus (A \cup B)$. In the language of elliptic equations, the decomposition above is the superposition principle for the Dirichlet problem in Ω . Since $\Omega \subset \mathbb{T}^d \setminus B$, we have $w_B = u_1 + u_2$ in Ω , where

$$-\mathcal{L}_\varepsilon u_1 = 1 \text{ in } \Omega, \quad u_1 = 0 \text{ on } \partial\Omega,$$

and

$$\mathcal{L}_\varepsilon u_2 = 0 \text{ in } \Omega, \quad u_2 = w_B \text{ on } \partial A, \quad u_2 = 0 \text{ on } \partial B.$$

By Dynkin's formula, as in Chapter 8, $u_1(w) = \mathbb{E}_w[\tau_{A \cup B}]$ and

$$u_2(w) = \int_{\partial A} w_B d\mathcal{H}_w, \quad \text{where } \mathcal{H}_w(E) = \mathbb{P}_w[w_{\tau_{A \cup B}} \in E]$$

is the harmonic measure of Ω at w for the operator $\mathcal{L}_\varepsilon$. In particular, $\mathcal{H}_w(\partial A) = q_{A,B}(w)$. Let T_ε^+ and T_ε^- be the maximum and the minimum of w_B on ∂A . Then $T_\varepsilon^- \leq T_\varepsilon \leq T_\varepsilon^+$, and

$$T_\varepsilon^- q_{A,B}(w) \leq w_B(w) - u_1(w) \leq T_\varepsilon^+ q_{A,B}(w).$$

By (a), $u_1 = o(T_\varepsilon)$ on D . By (b), the equilibrium potential satisfies $q_{A,B} = 1 - o(1)$ on D . By (c), $T_\varepsilon^+ - T_\varepsilon^- = o(T_\varepsilon)$. Hence

$$w_B(w) = T_\varepsilon (1 + o(1)) \quad \text{uniformly for } w \in D.$$

Estimates of the types (a) and (b) are established in [BEGK] for non-degenerate wells. Estimate (c) is the main obstacle. It asks that the oscillation of w_B on ∂A be small compared with its size. So (c) is a Harnack-type estimate for the equation of w_B along the well K_A , with constants uniform in ε . For a non-degenerate minimum, [BEGK] controls the oscillation on a ball of radius of order $\varepsilon^{\frac{1}{2}}$ by Harnack inequalities on balls of radius ε . When K_A is positive-dimensional, ∂A has diameter of order one. Then a chain of order ε^{-1} Harnack balls along K_A produces a constant exponential in ε^{-1} , which is useless against the polynomial factor $\varepsilon^{\lambda_A - \frac{d}{2}}$. The reason is the drift. In the variable $w = w_0 + rz$,

$$\mathcal{L}_\varepsilon = \frac{\varepsilon}{r^2} \left(\Delta_z - \frac{r}{\varepsilon} \nabla L \cdot \nabla_z \right).$$

So the Harnack constant on a ball of radius r is bounded independently of ε when $r \sup |\nabla L| \leq C\varepsilon$ on the ball, and each ball in a chain costs a fixed factor.

Problem 9.2 (Equilibration along a singular well). *Under the hypotheses of Theorem 9.4, prove (c), i.e., that $y \mapsto \mathbb{E}_y[\tau_B]$ is asymptotically constant on ∂A relative to T_ε .*

Heuristically, (c) holds because the process equilibrates within the well on a time scale polynomial in ε^{-1} , much shorter than T_ε . Along a flat stratum, the noise moves the process diffusively, covering distances of order one in time of order ε^{-1} . In spectral terms, a natural route to (c) is a lower bound, polynomial in ε , on the spectral gap of the dynamics reflected at the boundary of a neighborhood of K_A . This gap is the best constant in the Poincaré inequality for the normalized Gibbs measure there; see the remark on spectral gap and relaxation in Section 8.6. By the conjugation in Lemma 8.4, it is $\frac{1}{\varepsilon}$ times the smallest non-zero eigenvalue of the Witten Laplacian P_ε there, with the boundary condition conjugated from the Neumann condition. This is a question of the same type as Problem 10.1(b) in the next chapter. The potential $\frac{1}{4}|\nabla L|^2$ of P_ε vanishes on the whole well, and there is no barrier inside the well. So the gap should not be exponentially small. The limiting motion along the zero set of a real-analytic potential is studied by Larssonⁱ. We discuss the motion along the zero set in the next chapter.

ⁱM. Larsson, *Langevin dynamics along the zero set of real-analytic potentials*. arXiv:2608.09840.

CHAPTER 10

Entropic selection and the noise of training

Metastability concerns wells separated by barriers, which are crossed on time scales exponential in $\frac{1}{\varepsilon}$. But the optimal set of a network may contain strata of the same loss, joined inside W_0 with no barrier between them. Noisy dynamics should still migrate from a less flat stratum to a flatter one, driven purely by the volume transversal to the strata, on time scales polynomial in $\frac{1}{\varepsilon}$. This *entropic* drift toward flatness is a candidate mechanism for the simplicity bias of noisy training, i.e., its tendency to find simple solutions, where simplicity is measured by the learning coefficient as in Chapter 6. In this chapter, we discuss the smallest test case, MAIS-O75, and the effect of realistic noise.

10.1. Two strata of equal loss

On $\mathbb{T}^2$, consider

$$L(x, y) = \sin^2 x \sin^4 y.$$

As we saw in Chapter 6, its zero set W_0 is the union of the two vertical circles $\{x \in \{0, \pi\}\}$, where $\lambda = \frac{1}{2}$ off the crossings, and the two horizontal circles $\{y \in \{0, \pi\}\}$, where $\lambda = \frac{1}{4}$. At the four crossings, $\lambda = \frac{1}{4}$ and $m = 1$; see Figure 10.1.

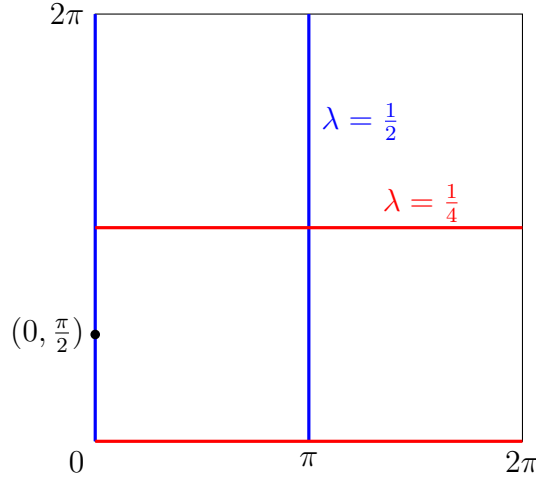


FIGURE 10.1. The zero set of $L(x, y) = \sin^2 x \sin^4 y$ on $\mathbb{T}^2$, with the vertical strata ($\lambda = \frac{1}{2}$) and the horizontal strata ($\lambda = \frac{1}{4}$)

Problem 10.1 (MAIS-O75). Let w_t be the Langevin dynamics with the potential $L(x, y) = \sin^2 x \sin^4 y$ at temperature ε , started at $(0, \frac{\pi}{2})$, the midpoint of a vertical stratum.

- (a) Let τ be the hitting time of $\{\text{dist}(w, \{y \in \{0, \pi\}\}) \leq \rho\}$ for a fixed small $\rho > 0$. Prove that $\mathbb{E}[\tau] = \varepsilon^{-b+o(1)}$, and determine b .
- (b) Prove that the inverse spectral gap of the generator is of order ε^{-a} up to logarithmic factors, and determine a .

(c) Replace the noise by the degenerate noise $\sqrt{2\varepsilon L(w_t)} dB_t$, which vanishes on W_0 , and start at $(\frac{\pi}{4}, \frac{\pi}{2}) \notin W_0$. Show that w_t converges almost surely to a point of W_0 , and determine the asymptotics of the distribution of the limit point as $\varepsilon \rightarrow 0$.

Problem 10.1 is a natural first target. A solution would give a rigorous instance of purely entropic selection between strata of a singular optimal set. The following heuristic explains the expected answer to (a).

Question 10.1. *Why should $b = 1$?*

Answer. Near the vertical stratum $\{x = 0\}$ and away from the crossings, $L(x, y) \approx x^2 \sin^4 y$. For a fixed y , the transversal coordinate x relaxes quickly to the Gaussian distribution with density proportional to $e^{-\frac{x^2 \sin^4 y}{\varepsilon}}$, whose variance is

$$\mathbb{E}[x^2] = \frac{\varepsilon}{2 \sin^4 y}.$$

The drift of the y -coordinate is

$$-\partial_y L = -4 \sin^2 x \sin^3 y \cos y \approx -4x^2 \sin^3 y \cos y,$$

and averaging over the transversal Gaussian gives the effective drift

$$-4 \cdot \frac{\varepsilon}{2 \sin^4 y} \cdot \sin^3 y \cos y = -2\varepsilon \cot y.$$

Equivalently, the transversal Gibbs mass at height y is

$$\int_{\mathbb{R}} e^{-\frac{x^2 \sin^4 y}{\varepsilon}} dx = \frac{\sqrt{\pi \varepsilon}}{\sin^2 y}.$$

So the y -coordinate feels the effective potential U , where

$$U(y) = -\varepsilon \log \frac{\sqrt{\pi \varepsilon}}{\sin^2 y} = 2\varepsilon \log |\sin y| + \text{const} \quad \text{and} \quad -U'(y) = -2\varepsilon \cot y.$$

This drift points toward the crossings at $y = 0$ and $y = \pi$, where the channel of nearly optimal parameters is widest. So the effective dynamics along the stratum is

$$dy_t = -2\varepsilon \cot y_t dt + \sqrt{2\varepsilon} dB_t.$$

As in the proof of Proposition 8.1, with the effective Gibbs density $e^{-\frac{U}{\varepsilon}}$, which is proportional to $\frac{1}{\sin^2 y}$, its generator is

$$\varepsilon \left(\partial_y^2 - 2 \cot y \partial_y \right) = \varepsilon \sin^2 y \partial_y \left(\frac{1}{\sin^2 y} \partial_y \right).$$

In the time $s = \varepsilon t$, the generator is divided by ε , and the equation becomes

$$dy = -2 \cot y ds + \sqrt{2} d\beta_s, \quad \text{where } \beta_s = \varepsilon^{\frac{1}{2}} B_{\frac{s}{\varepsilon}}$$

is again a standard Brownian motion, by the scaling of Brownian motion. This process reaches a neighborhood of $\{0, \pi\}$ in a time s of order one. Hence τ is of order ε^{-1} . The entropic drift is of the same order as the diffusion, so it biases the exit toward the crossings without changing the exponent.

The effective dynamics also predicts the constant. By Dynkin's formula, as in Section 8.3, the mean exit time $u(y)$ of the effective dynamics from $(\rho, \pi - \rho)$ solves

$$-\varepsilon (u'' - 2 \cot y u') = 1 \text{ on } (\rho, \pi - \rho) \quad \text{with } u(\rho) = u(\pi - \rho) = 0.$$

By the divergence form of the generator,

$$\left(\frac{u'}{\sin^2 y} \right)' = -\frac{1}{\varepsilon \sin^2 y}.$$

Since $u'(\frac{\pi}{2}) = 0$ by symmetry, integrating from $\frac{\pi}{2}$ gives

$$u'(y) = \frac{\sin y \cos y}{\varepsilon}.$$

Hence

$$u(y) = \frac{\sin^2 y - \sin^2 \rho}{2\varepsilon} \quad \text{and} \quad u\left(\frac{\pi}{2}\right) = \frac{\cos^2 \rho}{2\varepsilon}.$$

Without the entropic drift, the mean exit time of $\varepsilon \partial_y^2$ from the midpoint is

$$\frac{1}{2\varepsilon} \left(\frac{\pi}{2} - \rho \right)^2.$$

It is larger, since $\cos \rho = \sin(\frac{\pi}{2} - \rho) < \frac{\pi}{2} - \rho$. So the effective dynamics predicts

$$\mathbb{E}[\tau] = \frac{\cos^2 \rho}{2\varepsilon} (1 + o(1)),$$

i.e., $b = 1$ with an explicit constant.

Remark (The adiabatic reduction). In the language of Section 8.6, the answer is an adiabatic reduction of Born-Oppenheimerⁱ type. The transversal coordinate x is fast, and the coordinate y along the stratum is slow. For a fixed y , write $\sigma = \sin^4 y$. The Witten Laplacian in the variable x for the potential $x^2 \sigma$ is the harmonic oscillator

$$-\varepsilon^2 \partial_x^2 + \sigma^2 x^2 - \varepsilon \sigma,$$

whose lowest eigenvalue is $\varepsilon \sigma - \varepsilon \sigma = 0$ for every y . So, at order ε , the transversal energy produces no potential along the stratum. The effective dynamics comes from the dependence of the transversal ground state on y , i.e., from the transversal Gibbs mass computed above. The computation in the proof of Lemma 8.4, in dimension one and with U in place of L , turns the effective generator $\varepsilon \partial_y^2 - U'(y) \partial_y$ into

$$-\varepsilon^2 \partial_y^2 + \frac{1}{4} U'(y)^2 - \frac{\varepsilon}{2} U''(y) = \varepsilon^2 \left(-\partial_y^2 + 2 \cot^2 y + 1 \right),$$

since

$$U'(y) = 2\varepsilon \cot y \quad \text{and} \quad U''(y) = -\frac{2\varepsilon}{\sin^2 y}.$$

On $(\rho, \pi - \rho)$ with Dirichlet conditions, its eigenvalues are of order ε^2 . This corresponds to rates of order ε for the generator, in agreement with the heuristic value $b = 1$. The potential $2 \cot^2 y + 1$ grows like $\frac{2}{y^2}$ at the crossing $y = 0$, where the transversal frequency σ vanishes and the reduction breaks down.

Part (b) is a question of semiclassical spectral theory. By Section 8.6, it asks to show that the smallest non-zero eigenvalue of the Witten Laplacian P_ε of Lemma 8.4 on $\mathbb{T}^2$ is of order ε^{1+a} up to logarithmic factors. As explained there, the harmonic approximation is not available, since the leading part $\frac{1}{4} |\nabla L|^2$ of the potential of P_ε vanishes on the whole zero set W_0 , a union of four circles. On the exponential scale, the theorem of Holley-Kusuoka-Stroockⁱⁱ states the following.

ⁱM. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*. Ann. Physik 84 (1927), 457–484.

ⁱⁱR. Holley, S. Kusuoka, and D. Stroock, *Asymptotics of the spectral gap with applications to the theory of simulated annealing*. J. Funct. Anal. 83 (1989), 333–347.

For a smooth potential on a compact connected manifold, the spectral gap γ_ε of the generator satisfies

$$\lim_{\varepsilon \rightarrow 0} \varepsilon \log \gamma_\varepsilon = -h,$$

where h is the critical depth of the potential, roughly the largest height one has to climb from a local minimum to reach a global minimum. The critical points of L off W_0 are the four maxima $(\pm \frac{\pi}{2}, \pm \frac{\pi}{2})$. So every local minimum lies on the connected set W_0 of global minima, and $h = 0$. Hence the gap is not exponentially small, and (b) asks for the power of ε .

10.2. Degenerate noise

The noise of SGD is not isotropic, and it typically vanishes on the optimal set (Proposition 3.2 and the case of exact fits after it). Part (c) of Problem 10.1 replaces the Langevin noise by $\sqrt{2\varepsilon L} dB$, a caricature of minibatch noise vanishing at zero loss. The initial point is off the zero set, since on W_0 both the drift and the noise vanish, and the process stays at its initial point. For noise of this type, Wojtowytschⁱ proves the convergence to the optimal set and a preference for flatness provably different from the isotropic case. What remains open is quantitative, and it is part (c): how the selection is split between the learning coefficient and the basin of first arrival.

The dynamics of part (c) is

$$dw_t = -\nabla L(w_t) dt + \sqrt{2\varepsilon L(w_t)} dB_t.$$

By Itô's formula, as in Chapter 8, its noise of covariance $2\varepsilon L$ produces the second-order term $\varepsilon L \Delta$. So the generator is

$$\tilde{\mathcal{L}}_\varepsilon f = \varepsilon L \Delta f - \nabla L \cdot \nabla f = \varepsilon L^{1+\frac{1}{\varepsilon}} \operatorname{div} \left(L^{-\frac{1}{\varepsilon}} \nabla f \right).$$

As in the proof of Proposition 8.1, $\tilde{\mathcal{L}}_\varepsilon$ is symmetric with respect to the measure $L^{-1-\frac{1}{\varepsilon}} dw$ on $\mathbb{T}^2 \setminus W_0$, which plays the role of the Gibbs measure. By the volume law, this measure has infinite mass near every point of W_0 , since $1 + \frac{1}{\varepsilon}$ exceeds the learning coefficients $\frac{1}{2}$ and $\frac{1}{4}$. So it cannot be normalized, in accordance with the convergence of w_t to W_0 . Moreover,

$$\tilde{\mathcal{L}}_\varepsilon = L(\varepsilon \Delta - \nabla \log L \cdot \nabla).$$

Multiplying a generator by a positive function only changes the clock. Precisely, let $\hat{w}_s$ be the Langevin dynamics at temperature ε with the potential $\log L$, i.e.,

$$d\hat{w}_s = -\nabla \log L(\hat{w}_s) ds + \sqrt{2\varepsilon} d\beta_s \quad \text{with } \hat{w}_0 = w_0,$$

for a standard Brownian motion β . Then, before W_0 is reached, the process w_t has the same law as $\hat{w}_{s(t)}$, where the clock $s(t)$ solves

$$\frac{ds}{dt} = L(\hat{w}_{s(t)}) \quad \text{with } s(0) = 0.$$

Indeed, at $t = 0$, the derivative of $\mathbb{E}[f(\hat{w}_{s(t)})]$ is $L(w_0)$ times the generator of $\hat{w}_s$ applied to f at w_0 , which is $\tilde{\mathcal{L}}_\varepsilon f(w_0)$. So the two processes follow the same paths at different speeds, and the limit point in part (c) should be the point at which $\hat{w}_s$ reaches W_0 .

For $L(x, y) = \sin^2 x \sin^4 y$,

$$\log L(x, y) = 2 \log |\sin x| + 4 \log |\sin y|.$$

ⁱS. Wojtowytsch, *Stochastic gradient descent with noise of machine learning type. Part II: continuous time analysis*. J. Nonlinear Sci. 34 (2024), Paper No. 16.

So the two coordinates of $\hat{w}_s$ are independent one-dimensional diffusions,

$$dx_s = -2 \cot x_s ds + \sqrt{2\varepsilon} d\beta_s^1 \quad \text{and} \quad dy_s = -4 \cot y_s ds + \sqrt{2\varepsilon} d\beta_s^2.$$

Near $\{0, \pi\}$, their drifts point into $\{0, \pi\}$ and are of sizes $\frac{2}{\text{dist}(x, \{0, \pi\})}$ and $\frac{4}{\text{dist}(y, \{0, \pi\})}$. So the stratum on which $\hat{w}_s$ arrives is decided by the coordinate which reaches $\{0, \pi\}$ first. From $(\frac{\pi}{4}, \frac{\pi}{2})$, the coordinate y starts at the unstable equilibrium $\frac{\pi}{2}$ of its drift, while the noiseless flow of x reaches 0 at the time $s = \frac{1}{4} \log 2$. This suggests that, as $\varepsilon \rightarrow 0$, the limit point concentrates at $(0, \frac{\pi}{2})$, on a stratum with $\lambda = \frac{1}{2}$. So, from this initial point, the basin of first arrival decides, and not the learning coefficient.

CHAPTER 11

Superposition

In Section 4.3, we described the superposition hypothesis of Elhage et al. [EHO]: a network stores more features than its activation space has dimensions, as nearly orthogonal directions, and it tolerates the interference between them because only a few features are active at a time. The regular pentagon of Figure 4.1 stores five features in the plane. In this chapter, we state the hypothesis precisely and explain the mathematics behind it: room for exponentially many nearly orthogonal directions (concentration of measure on spheres), the price of near orthogonality (the Welch bound and spherical harmonics), and decoding (uncertainty principles and sparse recovery, where the sup norms of eigenfunctions enter). We end with the toy model of Elhage et al. and its open problems. For one active feature, unit norms, and zero biases, we prove that the regular pentagon is optimal.

Throughout this chapter, m is the dimension of the activation space and N is the number of features, usually with $N > m$. These letters have nothing to do with the multiplicity m of Chapter 6. Similarly, in this chapter, Φ and λ do not denote the target matrix and the learning coefficient of the other chapters. The inner product of $\mathbb{R}^m$ or $\mathbb{C}^m$ is $\langle \cdot, \cdot \rangle$, linear in the first variable. For unit vectors $v_1, \dots, v_N$, the *coherence* is

$$\mu = \max_{i \neq j} |\langle v_i, v_j \rangle|.$$

A matrix whose columns are unit vectors is called a *dictionary*, and its coherence is that of its columns. For $x \in \mathbb{R}^N$ or $\mathbb{C}^N$, $\|x\|_0$ is the number of nonzero entries of x , and x is called k -sparse if $\|x\|_0 \leq k$.

11.1. The hypothesis

Fix a trained network, a layer, and a probability distribution P on the inputs $z \in \mathcal{X}$. Write $h(z) \in \mathbb{R}^m$ for the activations of the layer at the input z .

Definition (Sparse linear model). Let $N, k \in \mathbb{N}$ and $\varepsilon \geq 0$. The activations admit an (N, k, ε) sparse linear model if there are unit vectors $v_1, \dots, v_N \in \mathbb{R}^m$, a vector $b \in \mathbb{R}^m$, and a measurable map $x : \mathcal{X} \rightarrow [0, \infty)^N$, the code, such that

$$\mathbb{E} \|x(z)\|_0 \leq k \quad \text{and} \quad \mathbb{E} \left\| h(z) - b - \sum_{i=1}^N x_i(z) v_i \right\|^2 \leq \varepsilon \mathbb{E} \|h(z) - \mathbb{E} h(z)\|^2,$$

where the expectations are over $z \sim P$.

The vector v_i is the direction of the i -th feature, and $x_i(z) \geq 0$ is its intensity. The *superposition hypothesis* says that the activations admit such a model with $N \gg m$, with directions that are only nearly orthogonal, and with $k \ll m$. The sparse autoencoders of Huben et al.ⁱ compute such codes for language models.

ⁱR. Huben, H. Cunningham, L. Smith, et al., *Sparse autoencoders find highly interpretable features in language models*. International Conference on Learning Representations, 2024.

Remark. Without further constraints, the definition has no content. Suppose that P is supported on finitely many inputs $z_1, \dots, z_M$ with $h(z_j) \neq 0$. Then

$$N = M, \quad b = 0, \quad v_j = \frac{h(z_j)}{\|h(z_j)\|}, \quad \text{and} \quad x(z_j) = \|h(z_j)\| e_j$$

give an $(M, 1, 0)$ sparse linear model. So the hypothesis needs either a bound on N in terms of the sample size, or a statistical model for the codes.

For the hypothesis to make sense, there must be room for $N \gg m$ nearly orthogonal directions (Sections 11.2 and 11.3), and a k -sparse code x must be recoverable from $\sum_i x_i v_i$ (Sections 11.4 and 11.5).

11.2. Room

Room comes from the concentration of measure on the sphere, through the following classical bound for spherical caps. Let ω be the normalized surface measure on $\mathbb{S}^{m-1}$.

Lemma 11.1. *For $v \in \mathbb{S}^{m-1}$ and $\varepsilon \in (0, \frac{1}{\sqrt{2}}]$, the cap*

$$C(v) = \{u \in \mathbb{S}^{m-1} : \langle u, v \rangle \geq \varepsilon\}$$

satisfies

$$\omega(C(v)) \leq e^{-\frac{\varepsilon^2 m}{2}}.$$

PROOF. Let $K = \{tu : u \in C(v), 0 \leq t \leq 1\}$ be the cone over the cap. By polar coordinates,

$$\text{Vol}(K) = \omega(C(v)) \text{Vol}(\mathbb{B}^m),$$

where $\mathbb{B}^m$ is the unit ball. For $u \in C(v)$ and $t \in [0, 1]$,

$$|tu - \varepsilon v|^2 = t^2 - 2t\varepsilon\langle u, v \rangle + \varepsilon^2 \leq t^2 - 2t\varepsilon^2 + \varepsilon^2.$$

The right side is a convex function of t , equal to $\varepsilon^2 \leq 1 - \varepsilon^2$ at $t = 0$ and to $1 - \varepsilon^2$ at $t = 1$. So K lies in the ball of radius $\sqrt{1 - \varepsilon^2}$ centered at εv , and

$$\omega(C(v)) \leq (1 - \varepsilon^2)^{\frac{m}{2}} \leq e^{-\frac{\varepsilon^2 m}{2}}. \quad \square$$

Theorem 11.2. *For $\varepsilon \in (0, \frac{1}{\sqrt{2}}]$, the sphere $\mathbb{S}^{m-1}$ contains at least $\frac{1}{2}e^{\frac{\varepsilon^2 m}{2}}$ unit vectors with pairwise $|\langle v_i, v_j \rangle| \leq \varepsilon$.*

PROOF. Two distinct unit vectors u, v with $|\langle u, v \rangle| \leq \varepsilon$ satisfy

$$|u - v|^2 = 2 - 2\langle u, v \rangle \geq 2 - 2\varepsilon > 0.$$

By compactness, such families have bounded cardinality. Let V be one of maximal cardinality. If $u \in \mathbb{S}^{m-1} \setminus V$, then $V \cup \{u\}$ is not admissible, so $|\langle u, v \rangle| > \varepsilon$ for some $v \in V$. This also holds for $u \in V$, with $v = u$. Hence

$$\mathbb{S}^{m-1} \subset \bigcup_{v \in V} (C(v) \cup C(-v)),$$

and Lemma 11.1 gives

$$1 \leq 2|V|e^{-\frac{\varepsilon^2 m}{2}}. \quad \square$$

For $m = 4096$ and $\varepsilon = 0.1$, Theorem 11.2 gives about 3.9×10^8 vectors. The same concentration of measure drives the equidistribution of random orthonormal bases of eigenfunctions in Section 11.5.

The exponent $\varepsilon^2 m$ is sharp up to a logarithm. Alonⁱ proved that if N unit vectors in $\mathbb{R}^m$ have pairwise $|\langle v_i, v_j \rangle| \leq \varepsilon$, where $N^{-\frac{1}{2}} \leq \varepsilon < \frac{1}{4}$, then

$$m \geq \frac{c \log N}{\varepsilon^2 \log \frac{1}{\varepsilon}},$$

where $c > 0$ is an absolute constant.

11.3. The price

Nearly orthogonal is not orthogonal. The following bound says how small the coherence of $N > m$ unit vectors can be.

Theorem 11.3 (Welch). *Let $v_1, \dots, v_N$ be unit vectors in $\mathbb{R}^m$ or $\mathbb{C}^m$, with $N \geq m$ and $N \geq 2$. Then*

$$\sum_{i,j=1}^N |\langle v_i, v_j \rangle|^2 \geq \frac{N^2}{m} \quad \text{and} \quad \mu^2 \geq \frac{N-m}{m(N-1)}.$$

Equality holds in the first inequality if and only if

$$\sum_{i=1}^N v_i v_i^* = \frac{N}{m} \mathbf{I}.$$

Equality holds in the second inequality if and only if, in addition, $|\langle v_i, v_j \rangle|$ is the same for all $i \neq j$.

The theorem is due to Welchⁱⁱ.

PROOF. Let $F = \sum_i v_i v_i^*$ be the frame operator, with eigenvalues $\lambda_1, \dots, \lambda_m \geq 0$. Then $\sum_j \lambda_j = \text{Tr } F = N$, and

$$\sum_{i,j} |\langle v_i, v_j \rangle|^2 = \text{Tr } (F^2) = \sum_{j=1}^m \lambda_j^2 \geq \frac{1}{m} \left(\sum_{j=1}^m \lambda_j \right)^2 = \frac{N^2}{m},$$

by the Cauchy-Schwarz inequality, with equality if and only if all λ_j are equal. For the second inequality, the N diagonal terms equal 1 and the $N(N-1)$ others are at most μ^2 , so

$$N + N(N-1)\mu^2 \geq \frac{N^2}{m}. \quad \square$$

A family with $\sum_i v_i v_i^* = \frac{N}{m} \mathbf{I}$ is a *tight frame*, that is,

$$\sum_{i=1}^N |\langle u, v_i \rangle|^2 = \frac{N}{m} \|u\|^2 \quad \text{for every } u.$$

The sum $\sum_{i,j} \langle v_i, v_j \rangle^2$ is the *frame potential*. By Benedetto-Fickusⁱⁱⁱ, for $N \geq m$, the minimizers of the frame potential over N unit vectors in $\mathbb{R}^m$ are exactly the tight frames, and every local

ⁱN. Alon, *Perturbed identity matrices have high rank: proof and applications*. Combin. Probab. Comput. 18 (2009), 3–15.

ⁱⁱL. Welch, *Lower bounds on the maximum cross correlation of signals*. IEEE Trans. Inform. Theory 20 (1974), 397–399.

ⁱⁱⁱJ. Benedetto and M. Fickus, *Finite normalized tight frames*. Adv. Comput. Math. 18 (2003), 357–385.

minimizer is a global minimizer. Five unit vectors in $\mathbb{R}^2$ form many tight frames, and the regular pentagon is one of them. This is the linear analogue of the pentagon problems of Section 11.7.

The linear programming method of Delsarte-Goethals-Seidelⁱ generalizes the first inequality of Theorem 11.3. Let P_ℓ be the Gegenbauer polynomial of degree ℓ for $\mathbb{S}^{m-1}$, normalized by $P_\ell(1) = 1$, and let $\mathcal{H}_\ell$ be the space of spherical harmonics of degree ℓ . The addition theorem

$$\sum_Y Y(x)Y(y) = \dim \mathcal{H}_\ell P_\ell(\langle x, y \rangle),$$

where Y runs over an orthonormal basis of $\mathcal{H}_\ell$ in $L^2(\omega)$, gives

$$\sum_{i,j=1}^N P_\ell(\langle v_i, v_j \rangle) = \frac{1}{\dim \mathcal{H}_\ell} \sum_Y \left| \sum_{i=1}^N Y(v_i) \right|^2 \geq 0. \quad (11.1)$$

So if $f = \sum_\ell f_\ell P_\ell$ with $f_\ell \geq 0$ for $\ell \geq 1$, then

$$\sum_{i,j=1}^N f(\langle v_i, v_j \rangle) \geq f_0 N^2.$$

For real unit vectors and $f(t) = t^2 = \frac{1}{m} + \frac{m-1}{m} P_2(t)$, this is the first inequality of Theorem 11.3. If, moreover, $f_0 > 0$ and $f(t) \leq 0$ for $-1 \leq t \leq \cos \theta$, then a family of unit vectors with pairwise angles at least θ satisfies

$$Nf(1) \geq \sum_{i,j=1}^N f(\langle v_i, v_j \rangle) \geq f_0 N^2.$$

So

$$N \leq \frac{f(1)}{f_0}.$$

On the circle $\mathbb{S}^1$, $P_\ell(\cos \theta) = \cos \ell \theta$, and (11.1) reads

$$\sum_{i,j=1}^N \cos \ell(\theta_i - \theta_j) = \left| \sum_{j=1}^N e^{i\ell\theta_j} \right|^2 \geq 0.$$

We use this in Section 11.7.

On $\mathbb{S}^2$, with the normalized measure ω , the space $\mathcal{H}_\ell$ has dimension $2\ell + 1$, and P_ℓ is the Legendre polynomial. The reproducing kernel of $\mathcal{H}_\ell$ is $(2\ell + 1)P_\ell(\langle x, y \rangle)$, and its normalized sections

$$e_x = (2\ell + 1)^{\frac{1}{2}} P_\ell(\langle x, \cdot \rangle) \quad \text{for } x \in \mathbb{S}^2$$

are the L^2 -normalized zonal harmonics. So

$$f(x) = \sqrt{2\ell + 1} \langle f, e_x \rangle \quad \text{for } f \in \mathcal{H}_\ell,$$

and e_x attains the maximal growth $\|e_x\|_\infty = \sqrt{2\ell + 1}$ of the sup norm. For $2 < p \leq 6$, the maximal growth of the L^p norm is attained by the Gaussian beams instead, and the authorⁱⁱ showed that a positive density subsequence of orthonormal spherical harmonics attains it.

ⁱP. Delsarte, J. Goethals, and J. Seidel, *Spherical codes and designs*. Geom. Dedicata 6 (1977), 363–388.

ⁱⁱX. Han, *Spherical harmonics with maximal L^p ($2 < p \leq 6$) norm growth*. J. Geom. Anal. 26 (2016), 378–398.

11.4. Decoding

Suppose that $y = \Phi x$ with an $m \times N$ dictionary Φ and a sparse x . When does y determine x , and how can we compute x ? The following theorem is due to Donoho-Eladⁱ and Gribonval-Nielsenⁱⁱ.

Theorem 11.4 (Donoho-Elad, Gribonval-Nielsen). *Let Φ be a real or complex $m \times N$ dictionary with coherence $\mu > 0$, and let x be k -sparse with*

$$k < \frac{1}{2} \left(1 + \frac{1}{\mu} \right).$$

Then x is the unique sparsest solution of $\Phi z = \Phi x$, and x is the unique minimizer of $\|z\|_1$ subject to $\Phi z = \Phi x$.

PROOF. Let $h \in \ker \Phi \setminus \{0\}$, and let $v_1, \dots, v_N$ be the columns of Φ . Taking the inner product of $\sum_l h_l v_l = 0$ with v_j gives

$$h_j = - \sum_{l \neq j} h_l \langle v_l, v_j \rangle.$$

So $|h_j| \leq \mu(\|h\|_1 - |h_j|)$, i.e.,

$$|h_j| \leq \frac{\mu}{1 + \mu} \|h\|_1 \quad \text{for every } j.$$

For a set J of indices with $|J| \leq k$, summing over $j \in J$ gives

$$\|h_J\|_1 \leq \frac{k\mu}{1 + \mu} \|h\|_1 < \frac{1}{2} \|h\|_1,$$

where h_J is the restriction of h to J . This is the null space property. Let $J = \text{supp } x$ and $z = x + h$. Then

$$\|z\|_1 = \|x_J + h_J\|_1 + \|h_{J^c}\|_1 \geq \|x\|_1 - \|h_J\|_1 + \|h_{J^c}\|_1 > \|x\|_1.$$

Suppose that $x' \neq x$ is another solution with $\|x'\|_0 \leq k$. Then $h = x' - x$ is supported on at most $2k$ indices, and summing the bound on $|h_j|$ over its support gives

$$\|h\|_1 \leq \frac{2k\mu}{1 + \mu} \|h\|_1 < \|h\|_1,$$

which is a contradiction. □

By Theorem 11.3, $\frac{1}{\mu} \lesssim \sqrt{m}$, so coherence alone guarantees recovery only for $k \lesssim \sqrt{m}$. In the Dirac-Fourier dictionary, this threshold comes from an uncertainty principle. On $\mathbb{Z}/N\mathbb{Z}$, the unitary Fourier transform is

$$\hat{f}(\xi) = N^{-\frac{1}{2}} \sum_{x \in \mathbb{Z}/N\mathbb{Z}} f(x) e^{-\frac{2\pi i x \xi}{N}}.$$

Theorem 11.5 (Donoho-Stark). *For $f : \mathbb{Z}/N\mathbb{Z} \rightarrow \mathbb{C}$ with $f \neq 0$,*

$$|\text{supp } f| \cdot |\text{supp } \hat{f}| \geq N, \quad \text{and hence} \quad |\text{supp } f| + |\text{supp } \hat{f}| \geq 2\sqrt{N}.$$

The theorem is due to Donoho-Starkⁱⁱⁱ. Equality holds when $N = n^2$ and f is the indicator of the subgroup $n\mathbb{Z}/N\mathbb{Z}$.

ⁱD. Donoho and M. Elad, *Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization*. Proc. Natl. Acad. Sci. USA 100 (2003), 2197–2202.

ⁱⁱR. Gribonval and M. Nielsen, *Sparse representations in unions of bases*. IEEE Trans. Inform. Theory 49 (2003), 3320–3325.

ⁱⁱⁱD. Donoho and P. Stark, *Uncertainty principles and signal recovery*. SIAM J. Appl. Math. 49 (1989), 906–931.

PROOF. By the Cauchy-Schwarz inequality and the Plancherel theorem,

$$\|\hat{f}\|_\infty \leq N^{-\frac{1}{2}} \|f\|_1 \leq \left(\frac{|\text{supp } f|}{N} \right)^{\frac{1}{2}} \|f\|_2 = \left(\frac{|\text{supp } f|}{N} \right)^{\frac{1}{2}} \|\hat{f}\|_2 \leq \left(\frac{|\text{supp } f| \cdot |\text{supp } \hat{f}|}{N} \right)^{\frac{1}{2}} \|\hat{f}\|_\infty.$$

Since $\hat{f} \neq 0$, the first inequality follows. The second follows from the inequality of arithmetic and geometric means. $\square$

For the dictionary $[I, \mathcal{F}^*] \in \mathbb{C}^{N \times 2N}$, where $\mathcal{F}$ is the matrix of the unitary Fourier transform, the coherence is $N^{-\frac{1}{2}}$. Theorem 11.5 shows that a representation with fewer than $\sqrt{N}$ terms is the unique sparsest one; see Donoho-Huoⁱ. For $N = p$ prime, Taoⁱⁱ proved the much stronger inequality

$$|\text{supp } f| + |\text{supp } \hat{f}| \geq p + 1,$$

which is sharp.

The toy model of Section 11.6 decodes in the simplest possible way: one matrix multiplication followed by a thresholding. For $W \in \mathbb{R}^{m \times N}$ with columns w_i , the largest positive interference is

$$\mu_+ = \max \left(0, \max_{i \neq j} \langle w_i, w_j \rangle \right).$$

Proposition 11.6. *Let $W \in \mathbb{R}^{m \times N}$ have unit columns, coherence μ , and largest positive interference μ_+ . Let $x \in [0, 1]^N$ have support J with $|J| \leq k$, and let $\beta \geq 0$. Let $\mathbf{1} = (1, \dots, 1)$, and let*

$$x' = \text{ReLU} \left(W^t W x - \beta \mathbf{1} \right).$$

Then

$$|x'_i - x_i| \leq (k-1)\mu + \beta \quad \text{for } i \in J, \quad \text{and} \quad x'_i \leq (k\mu_+ - \beta)_+ \quad \text{for } i \notin J.$$

PROOF. For every i ,

$$\left(W^t W x \right)_i = x_i \mathbf{1}_J(i) + \sum_{j \in J \setminus \{i\}} x_j \langle w_i, w_j \rangle.$$

For $i \in J$, the sum has at most $k-1$ terms of absolute value at most μ , and ReLU is 1-Lipschitz with $\text{ReLU}(x_i) = x_i$. For $i \notin J$, each term is at most $x_j \mu_+ \leq \mu_+$, and ReLU is nondecreasing. $\square$

So the error is at most the number of active features times the coherence, which is the quantitative content of “sparsity makes interference tolerable”. A negative bias $-\beta$ with $\beta \geq k\mu_+$ removes all false positives. Since the codes are nonnegative, the ReLU removes negative interference with no bias needed. This is why antipodal pairs, with $\langle w_i, w_j \rangle = -1$, cost little (Proposition 11.8).

11.5. Beyond the square-root bottleneck

Coherence-based guarantees stop at $k \sim \sqrt{m}$. Random dictionaries do much better, and the right notion is the following.

Definition (Restricted isometry). A matrix $A \in \mathbb{C}^{m \times N}$ has the restricted isometry property of order s with constant $\delta \in [0, 1)$ if

$$(1 - \delta) \|x\|^2 \leq \|Ax\|^2 \leq (1 + \delta) \|x\|^2 \quad \text{for every } s\text{-sparse } x \in \mathbb{C}^N.$$

The smallest such δ is denoted δ_s .

ⁱD. Donoho and X. Huo, *Uncertainty principles and ideal atomic decomposition*. IEEE Trans. Inform. Theory 47 (2001), 2845–2862.

ⁱⁱT. Tao, *An uncertainty principle for cyclic groups of prime order*. Math. Res. Lett. 12 (2005), 121–127.

Candèsⁱ proved that if $\delta_{2s} < \sqrt{2} - 1$, then every s -sparse x is the unique minimizer of $\|z\|_1$ subject to $Az = Ax$. Normalized Gaussian matrices with $m \geq C(\delta)s \log \frac{eN}{s}$ rows satisfy $\delta_s \leq \delta$ with high probability, by concentration, nets, and a union bound; see Baraniuk-Davenport-DeVore-Wakinⁱⁱ. So random dictionaries decode k -sparse codes up to $k \sim \frac{m}{\log(N/m)}$, far beyond $\sqrt{m}$. This is compressed sensing, founded by Candès-Romberg-Taoⁱⁱⁱ and Donoho^{iv}; see Foucart-Rauhut [FR] for a textbook account.

The mechanism behind the theorem of Candès is the null space property in the proof of Theorem 11.4. We prove the theorem with $\frac{1}{3}$ in place of $\sqrt{2} - 1$.

Proposition 11.7. *Let $A \in \mathbb{C}^{m \times N}$ satisfy $\delta_{2s} < \frac{1}{3}$. Then every $h \in \ker A \setminus \{0\}$ satisfies*

$$\|h_J\|_1 < \frac{1}{2}\|h\|_1 \quad \text{for every } J \text{ with } |J| \leq s.$$

Hence every s -sparse x is the unique minimizer of $\|z\|_1$ subject to $Az = Ax$.

PROOF. Let u, v be s -sparse with disjoint supports and $\|u\| = \|v\| = 1$. Then $u \pm v$ are $2s$ -sparse with $\|u \pm v\|^2 = 2$, so

$$\operatorname{Re}\langle Au, Av \rangle = \frac{1}{4} (\|A(u+v)\|^2 - \|A(u-v)\|^2) \leq \frac{1}{4} (2(1 + \delta_{2s}) - 2(1 - \delta_{2s})) = \delta_{2s}.$$

Replacing v by $-v$ and scaling, we obtain

$$|\operatorname{Re}\langle Au, Av \rangle| \leq \delta_{2s} \|u\| \|v\|$$

for all s -sparse u, v with disjoint supports. Order the indices so that $|h_1| \geq |h_2| \geq \dots$. Let J_0 be the first s indices, J_1 the next s , and so on. Every $|h_j|$ with $j \in J_k$ is at most the average of the $|h_i|$ with $i \in J_{k-1}$, so

$$\|h_{J_k}\| \leq \sqrt{s} \cdot \frac{\|h_{J_{k-1}}\|_1}{s} = s^{-\frac{1}{2}} \|h_{J_{k-1}}\|_1 \quad \text{for } k \geq 1.$$

Summing over k gives

$$\sum_{k \geq 1} \|h_{J_k}\| \leq s^{-\frac{1}{2}} \|h\|_1.$$

Since $Ah = 0$, we have

$$Ah_{J_0} = - \sum_{k \geq 1} Ah_{J_k}.$$

So

$$(1 - \delta_{2s}) \|h_{J_0}\|^2 \leq \|Ah_{J_0}\|^2 = - \sum_{k \geq 1} \operatorname{Re} \langle Ah_{J_0}, Ah_{J_k} \rangle \leq \delta_{2s} \|h_{J_0}\| s^{-\frac{1}{2}} \|h\|_1.$$

Hence

$$\|h_{J_0}\|_1 \leq \sqrt{s} \|h_{J_0}\| \leq \frac{\delta_{2s}}{1 - \delta_{2s}} \|h\|_1 < \frac{1}{2} \|h\|_1.$$

Since J_0 carries the largest entries, the same holds for every J with $|J| \leq s$. The last statement follows from this null space property as in the proof of Theorem 11.4. $\square$

ⁱE. Candès, *The restricted isometry property and its implications for compressed sensing*. C. R. Math. Acad. Sci. Paris 346 (2008), 589–592.

ⁱⁱR. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*. Constr. Approx. 28 (2008), 253–263.

ⁱⁱⁱE. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information*. IEEE Trans. Inform. Theory 52 (2006), 489–509.

^{iv}D. Donoho, *Compressed sensing*. IEEE Trans. Inform. Theory 52 (2006), 1289–1306.

Coherence controls the Gram matrix A^*A entry by entry, while the restricted isometry property controls all its principal $s \times s$ blocks at once: their eigenvalues lie in $[1 - \delta_s, 1 + \delta_s]$. For unit columns, the Gershgorin disc theorem gives $\delta_s \leq (s - 1)\mu$.

The harmonic analysis enters through structured random matrices. Choose m rows of the $N \times N$ unitary Fourier matrix of $\mathbb{Z}/N\mathbb{Z}$ independently and uniformly at random, and multiply by $(\frac{N}{m})^{\frac{1}{2}}$. Rudelson-Vershyninⁱ proved the restricted isometry property of order s with high probability for $m \gtrsim s \log N (\log s)^2 \log(s \log N)$. Bourgainⁱⁱ improved this to $m \gtrsim s \log s (\log N)^2$, and Haviv-Regevⁱⁱⁱ to $m \gtrsim s (\log s)^2 \log N$. In the other direction, consider the Walsh-Hadamard matrix of $(\mathbb{Z}/2\mathbb{Z})^n$, and suppose that $\min(s, \frac{N}{s}) > (\log N)^C$. Błasiok et al.^{iv} proved that, with high probability, the matrix formed by independent uniformly random rows has an s -sparse vector in its kernel unless the number of rows is at least of order $s \log s \log \frac{N}{s}$.

Question 11.1. *What is the smallest number of random rows of the Fourier matrix of $\mathbb{Z}/N\mathbb{Z}$ that guarantees the restricted isometry property of order s with high probability? Which powers of $\log s$ and $\log N$ are necessary?*

The problem is a relative of the $\Lambda(p)$ problem. Rudin^v called a set $E \subset \mathbb{Z}$ a $\Lambda(p)$ set, $p > 2$, if

$$\left\| \sum_{\xi \in E} a_\xi e^{i\xi t} \right\|_{L^p(\mathbb{T})} \leq C \|a\|_{\ell^2} \quad \text{for all coefficients } a.$$

Bourgain^{vi} proved that for every uniformly bounded orthonormal system $\varphi_1, \dots, \varphi_n$ and $p > 2$, a random subset of $\{1, \dots, n\}$ of size about $n^{\frac{2}{p}}$ satisfies such an inequality with a constant depending only on p and the bound of the system. Both the $\Lambda(p)$ theorem and the restricted isometry of partial Fourier matrices assert that a random part of a bounded orthonormal system satisfies a uniform norm inequality.

The restricted isometry theorem holds for every bounded orthonormal system. Let $\varphi_1, \dots, \varphi_N$ be orthonormal in $L^2(\nu)$ for a probability measure ν , with $\|\varphi_j\|_\infty \leq K$. Let $t_1, \dots, t_m$ be independent samples of ν , and let

$$A = m^{-\frac{1}{2}} (\varphi_j(t_l))_{l,j}.$$

If $m \geq CK^2 \delta^{-2} s (\log N)^4$, then $\delta_s \leq \delta$ with high probability; see Foucart-Rauhut [FR].

Remark (Restricted isometry as uniform sampling). For $x \in \mathbb{C}^N$, let $f_x = \sum_j x_j \varphi_j$, so that $\|f_x\|_{L^2(\nu)} = \|x\|$ and

$$\|Ax\|^2 = \frac{1}{m} \sum_{l=1}^m |f_x(t_l)|^2.$$

Hence

$$\delta_s = \sup \left\{ \left| \frac{1}{m} \sum_{l=1}^m |f_x(t_l)|^2 - 1 \right| : \|x\| = 1, \|x\|_0 \leq s \right\}.$$

ⁱM. Rudelson and R. Vershynin, *On sparse reconstruction from Fourier and Gaussian measurements*. Comm. Pure Appl. Math. 61 (2008), 1025–1045.

ⁱⁱJ. Bourgain, *An improved estimate in the restricted isometry problem*. Lecture Notes in Math. 2116 (2014), 65–70.

ⁱⁱⁱI. Haviv and O. Regev, *The restricted isometry property of subsampled Fourier matrices*. Lecture Notes in Math. 2169 (2017), 163–179.

^{iv}J. Błasiok, P. Lopatto, K. Luh, et al., *An improved lower bound for sparse reconstruction from subsampled Walsh matrices*. Discrete Anal. 2023, Paper No. 3.

^vW. Rudin, *Trigonometric series with gaps*. J. Math. Mech. 9 (1960), 203–227.

^{vi}J. Bourgain, *Bounded orthogonal systems and the $\Lambda(p)$ -set problem*. Acta Math. 162 (1989), 227–245.

So the restricted isometry property is a uniform law of large numbers: the empirical measure of the samples $t_1, \dots, t_m$ computes $\|f\|_{L^2(\nu)}^2$ with relative error at most δ_s , uniformly on the union of the $\binom{N}{s}$ coordinate subspaces of dimension s . For a fixed unit vector x with $\|x\|_0 \leq s$, the variables $|f_x(t_l)|^2$ are independent with mean 1, and, by the Cauchy-Schwarz inequality,

$$|f_x(t)|^2 \leq K^2 \|x\|_1^2 \leq sK^2.$$

This is where the sup norms enter. Bernstein's inequality for sums of independent bounded random variables, with the variance bound $\mathbb{E}|f_x(t_1)|^4 \leq sK^2$, gives

$$\mathbb{P} \left(\left| \frac{1}{m} \sum_{l=1}^m |f_x(t_l)|^2 - 1 \right| \geq \delta \right) \leq 2e^{-\frac{m\delta^2}{3sK^2}} \quad \text{for } 0 < \delta < 1.$$

A union bound over a net of $e^{C(\delta)s}$ points in the unit sphere of each coordinate subspace, and over the $\binom{N}{s} \leq (\frac{eN}{s})^s$ subspaces, as for the Gaussian matrices above, gives $\delta_s \leq \delta$ with high probability if $m \gtrsim K^2 \delta^{-2} s^2 \log N$. This loses a factor s . The known proofs replace the single union bound by union bounds at all dyadic scales (chaining), with the distance $\max_l |f_x(t_l) - f_{x'}(t_l)|$ between sparse unit vectors x and x' . The covering numbers for this distance are estimated by approximating x , which satisfies $\|x\|_1 \leq \sqrt{s}$, by averages of a few random multiples of coordinate vectors; see Foucart-Rauhut [FR, Chapter 12]. Each of these steps costs logarithms, and Question 11.1 asks which of them are necessary.

On the torus $\mathbb{T}^n$, the eigenfunctions $e^{ik \cdot t}$ of the Laplacian have $K = 1$, and we are back to random partial Fourier matrices. On the sphere $\mathbb{S}^2$, the spherical harmonics of degree at most L form a system of $N = (L+1)^2$ functions, and the zonal harmonics give $K^2 = 2L+1 \approx 2\sqrt{N}$. This is the sharpness of the sup norm bound of Soggeⁱ on the sphere. So the general theorem needs $m \gtrsim s\sqrt{N}(\log N)^4$ samples. By sampling from a different measure and preconditioning with a weight, Rauhut-Wardⁱⁱ reduced the number of samples to $m \gtrsim s(\log N)^4$ for Legendre polynomials, instead of $m \gtrsim sN(\log N)^4$, and to $m \gtrsim sN^{\frac{1}{4}}(\log N)^4$ for spherical harmonics. Burq-Dyatlov-Ward-Zworskiⁱⁱⁱ proved weighted L^∞ estimates for eigenfunctions on strictly convex surfaces of revolution by semiclassical analysis, and reached $m \gtrsim sN^{\frac{1}{6}}(\log N)^4$ on $\mathbb{S}^2$. So sup norms of eigenfunctions are exactly the input of this decoding theory.

Remark (Sampling measure and weight). The theorem on bounded orthonormal systems allows a change of the sampling measure. Let $\rho > 0$ be a probability density with respect to ν , and draw $t_1, \dots, t_m$ independently from $\rho\nu$. The functions $\psi_j = \rho^{-\frac{1}{2}}\varphi_j$ are orthonormal in $L^2(\rho\nu)$, and the weighted samples of $f = \sum_j x_j \varphi_j$ are

$$\rho(t_l)^{-\frac{1}{2}} f(t_l) = \sum_j x_j \psi_j(t_l).$$

So the theorem applies to the matrix $m^{-\frac{1}{2}}(\psi_j(t_l))_{l,j}$, with K replaced by the weighted sup norm

$$K_\rho = \max_j \sup_t \frac{|\varphi_j(t)|}{\sqrt{\rho(t)}}.$$

ⁱC. Sogge, *Concerning the L^p norm of spectral clusters for second-order elliptic operators on compact manifolds*. J. Funct. Anal. 77 (1988), 123–138.

ⁱⁱH. Rauhut and R. Ward, *Sparse Legendre expansions via ℓ_1 -minimization*. J. Approx. Theory 164 (2012), 517–533; *Sparse recovery for spherical harmonic expansions*. arXiv:1102.4097.

ⁱⁱⁱN. Burq, S. Dyatlov, R. Ward, and M. Zworski, *Weighted eigenfunction estimates with applications to compressed sensing*. SIAM J. Math. Anal. 44 (2012), 3481–3501.

For example, the functions $\varphi_\ell = \sqrt{2\ell+1}P_\ell$, $0 \leq \ell < N$, are orthonormal for $\nu = \frac{1}{2}dx$ on $[-1, 1]$, and $K = \sqrt{2N-1}$ is attained at $x = \pm 1$. Take the Chebyshev density

$$\rho(x) = \frac{2}{\pi} (1-x^2)^{-\frac{1}{2}}.$$

By the inequality of Bernstein (see Szegőⁱ),

$$|P_\ell(\cos \theta)| < \left(\frac{2}{\pi \ell \sin \theta} \right)^{\frac{1}{2}} \quad \text{for } \ell \geq 1 \text{ and } 0 < \theta < \pi.$$

So for $\ell \geq 1$ and $x = \cos \theta$,

$$\frac{|\varphi_\ell(x)|}{\sqrt{\rho(x)}} < \sqrt{2\ell+1} \left(\frac{2}{\pi \ell \sin \theta} \right)^{\frac{1}{2}} \left(\frac{\pi}{2} \sin \theta \right)^{\frac{1}{2}} = \left(\frac{2\ell+1}{\ell} \right)^{\frac{1}{2}} \leq \sqrt{3},$$

and $\rho^{-\frac{1}{2}} \leq \sqrt{\frac{\pi}{2}}$ handles $\ell = 0$. So $K_\rho \leq \sqrt{3}$ for every N , and $m \gtrsim s(\log N)^4$ samples from the Chebyshev measure suffice. This is the result of Rauhut-Ward for Legendre polynomials. On $\mathbb{S}^2$, the weighted eigenfunction estimates of Burq-Dyatlov-Ward-Zworski are bounds for K_ρ , and Question 11.2 asks how far the choice of ρ can reduce the number of samples.

Question 11.2. *What is the smallest exponent $\alpha \geq 0$ such that every s -sparse expansion in the first N standard spherical harmonics $Y_\ell^k(\phi, \theta) = C_{\ell,k} P_\ell^k(\cos \phi) e^{ik\theta}$ on $\mathbb{S}^2$ can be recovered by ℓ^1 minimization from $m \lesssim sN^\alpha (\log N)^C$ samples, for some choice of sampling measure and weight?*

The obstruction is the standard basis: the zonal harmonics are large at the poles. A random orthonormal basis of each $\mathcal{H}_\ell$ behaves much better. Burq-Lebeauⁱⁱ proved that, on the spheres, almost every orthonormal basis of L^2 made of spherical harmonics has all its elements uniformly bounded in L^p for every $p < \infty$, with L^∞ norms of order at most $\sqrt{\log \ell}$. The authorⁱⁱⁱ proved that random orthonormal bases of eigenfunctions are almost surely equidistributed at small scales on compact manifolds whose isometry group acts transitively and whose multiplicities grow at least logarithmically, e.g., on the spheres and on the tori $\mathbb{T}^n$, $n \geq 5$. Both results rest on the concentration of measure on the high-dimensional spheres of the eigenspaces, as in Lemma 11.1.

For sparse recovery, a version with high probability suffices, and it follows from Lemma 11.1 directly. Let u be uniformly distributed on the unit sphere of $\mathcal{H}_\ell$ in $L^2(\omega)$, of dimension $d = 2\ell + 1$. Each element of a random orthonormal basis of $\mathcal{H}_\ell$ is distributed in this way. For $x \in \mathbb{S}^2$, we have $u(x) = \sqrt{d} \langle u, e_x \rangle$ with the zonal harmonic e_x above. So Lemma 11.1, applied to the caps around $\pm e_x$, gives

$$\mathbb{P}(|u(x)| \geq t) \leq 2e^{-\frac{t^2}{2}} \quad \text{for } 0 < t \leq \sqrt{\frac{d}{2}}.$$

The restriction of u to a great circle is a trigonometric polynomial of degree at most ℓ , so Bernstein's inequality for trigonometric polynomials bounds the gradient of u by $\ell \|u\|_\infty$. Hence $\|u\|_\infty \leq 2 \max_{x \in X} |u(x)|$ for every set $X \subset \mathbb{S}^2$ that meets each geodesic ball of radius $\frac{1}{2\ell}$, and such a set with $O(\ell^2)$ points exists. A union bound over X and over the $N = (L+1)^2$ basis elements of degree at most L shows that, with probability at least $1 - \frac{C}{N}$, all of them satisfy

$$\|u\|_\infty \leq C \sqrt{\log N}.$$

ⁱG. Szegő, *Orthogonal Polynomials*. American Mathematical Society, 1975.

ⁱⁱN. Burq and G. Lebeau, *Injectons de Sobolev probabilistes et applications*. Ann. Sci. Éc. Norm. Supér. (4) 46 (2013), 917–962.

ⁱⁱⁱX. Han, *Small scale equidistribution of random eigenbases*. Comm. Math. Phys. 349 (2017), 425–440.

For $\ell \lesssim \log N$, this follows from the trivial bound $\|u\|_\infty \leq \sqrt{d}$. Since the samples are independent of the basis, the theorem on bounded orthonormal systems with $K^2 \lesssim \log N$ shows that $m \gtrsim s(\log N)^5$ independent uniform samples give the restricted isometry property of order s with high probability. So for sparse recovery, a random basis of spherical harmonics behaves like the Fourier basis of the torus, up to one logarithm.

Question 11.3. *Choose a random orthonormal basis of $\mathcal{H}_\ell$ independently for each $\ell \leq L$, and let $N = (L + 1)^2$. What is the smallest power C such that $m \lesssim s(\log N)^C$ independent uniform samples on $\mathbb{S}^2$ give the restricted isometry property of order s for the resulting system with high probability? Is there, for each ℓ , an orthonormal basis of $\mathcal{H}_\ell$ whose elements are bounded in L^∞ by a constant independent of ℓ ? Is there an explicit one whose elements are bounded by a power of $\log \ell$?*

The second part of Question 11.3 is a classical question on uniformly bounded orthonormal bases of eigenfunctions on $\mathbb{S}^2$. By the argument above, bases bounded by $C\sqrt{\log N}$ exist. A uniformly bounded basis would give $K \lesssim 1$ and $m \lesssim s(\log N)^4$, and an explicit basis would give an explicit sampling system.

11.6. The toy model

The toy model of Elhage et al. (Section 4.3) is the smallest network in which superposition can be studied. The data $x \in \mathbb{R}^N$ has independent coordinates. Each x_i equals 0 with probability $S \in (0, 1)$, the sparsity, and is otherwise uniform on $[0, 1]$. Feature i carries an importance $I_i > 0$. The network compresses x to $Wx \in \mathbb{R}^m$ and reconstructs

$$x' = \text{ReLU}(W^t W x + b) \quad \text{with } W \in \mathbb{R}^{m \times N} \text{ and } b \in \mathbb{R}^N,$$

and the loss is

$$L(W, b) = \mathbb{E} \sum_{i=1}^N I_i (x_i - x'_i)^2.$$

The column w_i of W is the direction in which feature i is stored. Elhage et al. also consider the linear model $x' = W^t W x + b$.

Elhage et al. trained these models, e.g., with $N = 20$, $m = 5$, and $I_i = 0.7^i$. They observed that as S increases, the ReLU model passes from storing the m most important features orthogonally to storing more than m features in superposition, and that the superposed features form antipodal pairs, triangles, pentagons, tetrahedra, and square antiprisms. For $N = 2$ and $m = 1$, they compared candidate configurations in closed form. On the side of theorems, Chen et al.ⁱ derived a closed formula for the loss, found that for $m = 2$ the regular k -gons are critical points, and related their local learning coefficients (Chapter 6) to phase transitions. Ivanov et al.ⁱⁱ proved that the optimal configurations are organized by the eigenspaces of the frame operator WW^t , under a hypothesis called capacity saturation, and Cowsik-Dolev-Infangerⁱⁱⁱ computed the loss at high sparsity for a permutation-symmetric ansatz. Global minimality of the observed polytopes, e.g., of the pentagon for $(N, m) = (5, 2)$, is open. For one active feature, unit norms, and zero biases, Proposition 11.9 proves that the regular pentagon is optimal.

The ReLU is necessary for superposition. For the linear model, suppose that the coordinates of x are independent with a common variance and that $I_1 \geq \dots \geq I_N$ with $I_m > I_{m+1}$. Then

ⁱZ. Chen, E. Lau, J. Mendel, et al., *Dynamical versus Bayesian phase transitions in a toy model of superposition*. arXiv:2310.06301.

ⁱⁱG. Ivanov, N. Oozeer, S. Raval, et al., *Spectral superposition: a theory of feature geometry*. arXiv:2602.02224.

ⁱⁱⁱA. Cowsik, K. Dolev, and A. Infanger, *The Persian rug: solving toy models of superposition using large-scale symmetries*. arXiv:2410.12101.

every minimizer stores the m most important features in orthonormal directions and drops the others, by the theorem of Eckart-Youngⁱ. In the network form, this is the theorem of Baldi-Hornik (Section 3.5) that linear autoencoders perform principal component analysis. For the data of the toy model, each coordinate has the variance

$$\mathbb{E}x_i^2 - (\mathbb{E}x_i)^2 = \frac{1-S}{3} - \frac{(1-S)^2}{4} = \frac{(1-S)(1+3S)}{12} =: v(S).$$

The number $v(S)$ is the cost of a dropped feature in the ReLU model. Indeed, if $w_i = 0$, then $x'_i = \text{ReLU}(b_i)$ is a constant, and

$$\mathbb{E}(x_i - c)^2 \geq v(S) \quad \text{for every constant } c,$$

with equality for $c = \mathbb{E}x_i = \frac{1-S}{2}$.

Proposition 11.8. *Let $I_1 = \dots = I_N = 1$.*

- (1) *Orthogonal storage: The infimum of L over the (W, b) whose nonzero columns are pairwise orthogonal equals $(N - m)v(S)$.*
- (2) *Antipodal storage: If $N \geq 2m$, the configuration $w_{2l-1} = e_l$, $w_{2l} = -e_l$ for $l \leq m$, $w_i = 0$ for $i > 2m$, with $b_i = 0$ for $i \leq 2m$ and $b_i = \frac{1-S}{2}$ for $i > 2m$, has loss*

$$L = \frac{m(1-S)^2}{3} + (N - 2m)v(S).$$

In particular, the loss in (2) is smaller than the infimum in (1) if and only if $S > \frac{3}{7}$. For $N = 2$ and $m = 1$, the antipodal pair $W = (1, -1)$, $b = 0$ has loss $\frac{(1-S)^2}{3}$. The configuration $W = (1, 0)$, which stores feature 1 and drops feature 2, has loss $v(S)$ with the optimal bias $b = (0, \frac{1-S}{2})$. So the pair is better if and only if $S > \frac{3}{7}$.

PROOF. (1) At most m columns are nonzero, and each zero column costs at least $v(S)$, so $L \geq (N - m)v(S)$. The configuration $W = [I_m \ 0]$ with $b_i = 0$ for $i \leq m$ and $b_i = \frac{1-S}{2}$ for $i > m$ reconstructs x_i exactly for $i \leq m$ and has loss $(N - m)v(S)$.

(2) For $i = 2l - 1$ with partner $j = 2l$, we have $(W^t W x)_i = x_i - x_j$, so $x'_i = \text{ReLU}(x_i - x_j)$. If $x_j = 0$, the error is 0. If $x_i = 0$, the output is $\text{ReLU}(-x_j) = 0$, and the error is 0. If both are positive, the error is $x_i - \text{ReLU}(x_i - x_j) = \min(x_i, x_j)$. Both are positive with probability $(1 - S)^2$, and for independent uniform U, U' ,

$$\mathbb{E} \min(U, U')^2 = \int_0^1 t^2 \cdot 2(1 - t) dt = \frac{1}{6},$$

since $\min(U, U')$ has density $2(1 - t)$ on $[0, 1]$. So each of the $2m$ paired coordinates costs $\frac{(1-S)^2}{6}$, and each dropped one costs $v(S)$. Finally, antipodal storage is better if and only if

$$\frac{m(1-S)^2}{3} < mv(S),$$

i.e., $4(1 - S) < 1 + 3S$, i.e., $S > \frac{3}{7}$. □

So for $N \geq 2m$ and $S > \frac{3}{7}$, the infimum of L is strictly below the infimum over orthogonal configurations, and superposition is forced. Note that the antipodal pair has coherence 1. The ReLU filters the negative interference (Proposition 11.6).

ⁱC. Eckart and G. Young, *The approximation of one matrix by another of lower rank*. Psychometrika 1 (1936), 211–218.

The geometry of the toy model is a problem of minimizing an energy of points on a sphere,

$$\min_{x_1, \dots, x_N \in \mathbb{S}^{m-1}} \sum_{i \neq j} f(\langle x_i, x_j \rangle).$$

Cohn-Kumarⁱ proved that certain configurations, among them the regular polygons on $\mathbb{S}^1$, minimize the energy simultaneously for all f that are completely monotone functions of the squared distance $2 - 2t$. These configurations are called *universally optimal*. For $f(t) = t^2$, the minimizers are the tight frames, by Benedetto-Fickus. For the toy model with one active feature, unit norms, and zero biases, the energy is $f(t) = \text{ReLU}(t)^2$ (Section 11.7). As a function of $r = 2 - 2t$, it equals $(1 - \frac{r}{2})_+^2$, which vanishes for $r \geq 2$ without vanishing identically. Since completely monotone functions are real-analytic, it is not completely monotone, and the theorem of Cohn-Kumar does not apply.

11.7. Problems

In this section, S is the sparsity and $p = 1 - S$ is the probability that a feature is active. The importances are equal. We first consider $(N, m) = (5, 2)$, where Elhage et al. observed the regular pentagon (Figure 4.1). Let $L_1(W, b)$ be the loss $\mathbb{E} \|x - \text{ReLU}(W^t W x + b)\|^2$ of the toy model, with x conditioned to have exactly one nonzero coordinate, chosen uniformly among the five, with value uniform on $[0, 1]$. Explicitly,

$$L_1(W, b) = \frac{1}{5} \sum_{i=1}^5 \left(\mathbb{E}_u \left(u - \text{ReLU} \left(u \|w_i\|^2 + b_i \right) \right)^2 + \sum_{j \neq i} \mathbb{E}_u \text{ReLU} \left(u \langle w_i, w_j \rangle + b_j \right)^2 \right),$$

where u is uniform on $[0, 1]$.

Problem 11.1 (MAIS-O42). *Prove that L_1 attains its infimum over $W \in \mathbb{R}^{2 \times 5}$ and $b \in \mathbb{R}^5$, and that every global minimizer is a regular pentagon, that is, its columns have a common norm $r > 0$, and for some permutation τ of $\{1, \dots, 5\}$,*

$$\langle w_{\tau(i)}, w_{\tau(j)} \rangle = r^2 \cos \frac{2\pi(i-j)}{5} \quad \text{for all } i, j.$$

For $b = 0$,

$$L_1(W, 0) = \frac{1}{15} \left(\sum_{i=1}^5 \left(1 - \|w_i\|^2 \right)^2 + \sum_{i \neq j} \text{ReLU} \left(\langle w_i, w_j \rangle \right)^2 \right).$$

The linear analogue $\frac{1}{15} \|I - W^t W\|_F^2$, where $\|\cdot\|_F$ is the Frobenius norm, is minimized by every W with $W^t W$ an orthogonal projection of rank 2, i.e., by all the tight frames with $\sum_i w_i w_i^t = I$. The ReLU must select the pentagon among them. For unit columns $w_j = (\cos \theta_j, \sin \theta_j)$ and $b = 0$, the problem becomes the minimization of the energy

$$E(\theta) = \sum_{i \neq j} h(\theta_i - \theta_j) \quad \text{with } h(\theta) = (\cos \theta)_+^2$$

on the circle. The regular pentagon has

$$E = 10 \cos^2 \frac{2\pi}{5} \approx 0.954915,$$

since each point has two neighbors at the angle $\frac{2\pi}{5}$ and the other angles have negative cosine.

ⁱH. Cohn and A. Kumar, *Universally optimal distribution of points on spheres*. J. Amer. Math. Soc. 20 (2007), 99–148.

The linear programming method of Section 11.3 gives lower bounds for E . Let

$$g(\theta) = a_0 + \sum_{k=1}^K a_k \cos k\theta \quad \text{with } a_k \geq 0 \text{ for } k \geq 1 \quad \text{and} \quad g \leq h \text{ on } \mathbb{R}.$$

Then for every $\theta_1, \dots, \theta_5$,

$$E(\theta) \geq \sum_{i \neq j} g(\theta_i - \theta_j) = \sum_{i,j} g(\theta_i - \theta_j) - 5g(0) = 25a_0 + \sum_{k=1}^K a_k \left| \sum_{j=1}^5 e^{ik\theta_j} \right|^2 - 5g(0) \geq 20a_0 - 5 \sum_{k=1}^K a_k.$$

So every admissible g gives the lower bound $20a_0 - 5 \sum_{k \geq 1} a_k$, and the best one is the value of a linear program in $(a_0, \dots, a_K)$.

The linear program has a dual, which tells where g must touch h . A configuration θ defines the positive measure $\rho_\theta = \sum_{i \neq j} \delta_{\theta_i - \theta_j}$ on $\mathbb{R}/2\pi\mathbb{Z}$, and

$$\int d\rho_\theta = 20, \quad \int \cos k\phi d\rho_\theta(\phi) = \left| \sum_{j=1}^5 e^{ik\theta_j} \right|^2 - 5 \geq -5, \quad \text{and} \quad E(\theta) = \int h d\rho_\theta.$$

The minimization of $\int h d\rho$ over the positive measures ρ with these constraints for $1 \leq k \leq K$ is a linear program in ρ , dual to the program in $(a_0, \dots, a_K)$. For admissible g and ρ ,

$$\int h d\rho \geq \int g d\rho = 20a_0 + \sum_{k=1}^K a_k \int \cos k\phi d\rho(\phi) \geq 20a_0 - 5 \sum_{k=1}^K a_k.$$

For the regular pentagon θ^* , we have $\sum_j e^{ik\theta_j^*} = 0$ if and only if $5 \nmid k$. So both inequalities are equalities for $\rho = \rho_{\theta^*}$ if and only if $g = h$ on the support $\{\pm \frac{2\pi}{5}, \pm \frac{4\pi}{5}\}$ of ρ_{θ^*} , and $a_k = 0$ whenever $5 \mid k$. This is complementary slackness. Since $g \leq h$, these contacts force $g' = h'$ at $\frac{2\pi}{5}$, where h is smooth, and $g' = 0$ at $\frac{4\pi}{5}$, where h vanishes nearby. A numerical solution with $K = 7$ uses only the frequencies 1, 2, 3, 6, 7, and it also touches h at 0 and π . These six contact conditions form a linear system for $(a_0, a_1, a_2, a_3, a_6, a_7)$, and its solution is an exact certificate; see Figure 11.1.

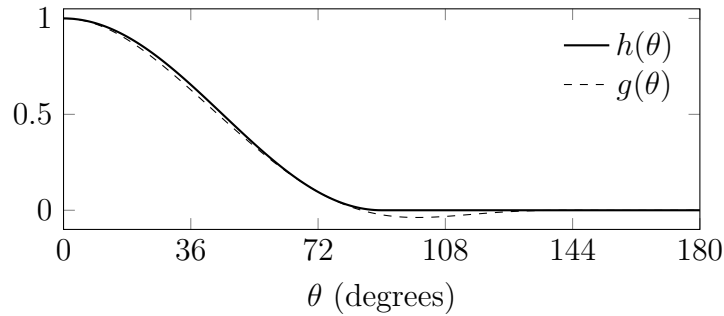


FIGURE 11.1. The ReLU energy $h(\theta) = (\cos \theta)_+^2$ and the certificate g of Proposition 11.9; $g \leq h$, with contact at 0° , 72° , 144° , and 180°

Proposition 11.9. *For all $\theta_1, \dots, \theta_5 \in \mathbb{R}$,*

$$E(\theta) \geq 10 \cos^2 \frac{2\pi}{5} = \frac{15 - 5\sqrt{5}}{4},$$

with equality if and only if $e^{i\theta_1}, \dots, e^{i\theta_5}$ are the vertices of a regular pentagon.

PROOF. Let

$$\begin{aligned} a_0 &= \frac{7 - \sqrt{5}}{20}, & a_1 &= \frac{6 + 2\sqrt{5}}{25}, & a_2 &= \frac{19 + 3\sqrt{5}}{100}, \\ a_3 &= \frac{17 - 4\sqrt{5}}{100}, & a_6 &= \frac{\sqrt{5} - 2}{50}, & a_7 &= \frac{9 - 4\sqrt{5}}{100}. \end{aligned}$$

They are positive. Let $g(\theta) = a_0 + \sum_k a_k \cos k\theta$, where k runs over $\{1, 2, 3, 6, 7\}$. Then $g(0) = 1$, and

$$20a_0 - 5 \sum_k a_k = 25a_0 - 5g(0) = \frac{15 - 5\sqrt{5}}{4}.$$

We show that $g \leq h$. Write $t = \cos \theta$. Then $\cos k\theta = T_k(t)$, where T_k is the Chebyshev polynomial of degree k , so g is a polynomial in t of degree 7 with leading coefficient $2^6 a_7$. Let

$$c_1 = \cos \frac{2\pi}{5} = \frac{\sqrt{5} - 1}{4} \quad \text{and} \quad c_2 = \cos \frac{4\pi}{5} = -\frac{\sqrt{5} + 1}{4}.$$

A direct computation gives

$$\begin{aligned} t^2 - g &= 64a_7(1 - t)(t - c_1)^2 Q_1(t), \\ -g &= 64a_7(1 + t)(t - c_2)^2 Q_2(t), \end{aligned}$$

where

$$\begin{aligned} Q_1(t) &= t^4 + \frac{5 + 3\sqrt{5}}{2}t^3 + \frac{27 + 13\sqrt{5}}{8}t^2 + \frac{5 + \sqrt{5}}{8}t + \frac{55 + 25\sqrt{5}}{16}, \\ Q_2(t) &= -t^4 - \frac{1 + \sqrt{5}}{2}t^3 + \frac{37 + 13\sqrt{5}}{8}t^2 - \frac{45 + 19\sqrt{5}}{8}t + \frac{5 + 5\sqrt{5}}{16}. \end{aligned}$$

The coefficients of Q_1 are positive, so $Q_1 > 0$ on $[0, 1]$. For $t \in [-1, 0]$, we have $t^4 \leq t^2$, so

$$-t^4 + \frac{37 + 13\sqrt{5}}{8}t^2 \geq 0,$$

and the other terms of $Q_2(t)$ are nonnegative, with a positive constant term. So $Q_2 > 0$ on $[-1, 0]$. Hence $g \leq t^2 = h$ for $t \in [0, 1]$, and $g \leq 0 = h$ for $t \in [-1, 0]$, with equality only for $t \in \{1, c_1, c_2, -1\}$, i.e., for $\theta \in D$, where

$$D = \left\{0, \pm \frac{2\pi}{5}, \pm \frac{4\pi}{5}, \pi\right\} \quad \text{modulo } 2\pi.$$

By the lower bound for E above,

$$E(\theta) \geq 20a_0 - 5 \sum_k a_k = \frac{15 - 5\sqrt{5}}{4},$$

and equality holds if and only if $\sum_j e^{ik\theta_j} = 0$ for $k \in \{1, 2, 3, 6, 7\}$ and $\theta_i - \theta_j \in D$ for all $i \neq j$.

The regular pentagon satisfies these conditions. Conversely, suppose that they hold. If some difference, say $\theta_2 - \theta_1$, equals π modulo 2π , then every $\theta_j - \theta_1$ lies in $D \cap (D + \pi) = \{0, \pi\}$ modulo 2π . So $\sum_j e^{i\theta_j}$ is an odd multiple of $e^{i\theta_1}$, which contradicts $\sum_j e^{i\theta_j} = 0$. Hence all the θ_j lie in $\theta_1 + \frac{2\pi}{5}\mathbb{Z}$. For $l \in \mathbb{Z}/5\mathbb{Z}$, let n_l be the number of j with $\theta_j \equiv \theta_1 + \frac{2\pi l}{5}$ modulo 2π . Then

$$\sum_{l \in \mathbb{Z}/5\mathbb{Z}} n_l e^{\frac{2\pi i k l}{5}} = 0$$

for $k = 1, 2$, and, by complex conjugation, for $k = 3, 4$. So the discrete Fourier transform of $(n_0, \dots, n_4)$ vanishes at all nonzero frequencies, all n_l are equal, and $n_l = 1$. $\square$

So the regular pentagon minimizes L_1 among the configurations with unit columns and zero biases, uniquely up to rotations and relabeling. The case of free norms and biases in Problem 11.1 remains open. The pentagon is special: for 7 points, numerical minimization finds configurations with coincident points whose energy is below that of the regular heptagon ($5.4320 < 5.4424$). So the energy E is not minimized by regular polygons in general.

Problem 11.2 (MAIS-O40). *Let $(N, m) = (5, 2)$, and let $L(W, b) = \mathbb{E}\|x - \text{ReLU}(W^t W x + b)\|^2$ with the data of the toy model. Prove that there is a nonempty open interval $J \subset (0, 1)$ such that for every $p \in J$, L attains its infimum, and every global minimizer is a regular pentagon in the sense of Problem 11.1.*

Let $G = O(m) \times \Sigma_N$ act by $(W, b) \mapsto (QWP^{-1}, Pb)$, for Q orthogonal and P a permutation matrix. Consider the penalty

$$R(W) = \sum_{i \neq j} \langle w_i, w_j \rangle^2.$$

The loss L and the penalty R are G -invariant. Call $L + \lambda R$ identifiable at (N, m, S, λ) if its set of global minimizers is nonempty and is a single G -orbit.

- Problem 11.3** (MAIS-O47). (1) *For $(N, m) = (5, 2)$ and $\lambda = 0$, prove that there is $S_0 < 1$ such that L is identifiable for every $S \in (S_0, 1)$, the orbit being the regular pentagon with its optimal norms and biases, or prove that no such S_0 exists.*
 (2) *Find (N, m, S, λ) with $\lambda > 0$ such that $L + \lambda R$ is identifiable while L is not, or prove that the penalty never creates identifiability.*

The three problems are linked as $p \rightarrow 0$. Splitting according to the number of active features,

$$L(W, b) = (1 - p)^5 \sum_{i=1}^5 \text{ReLU}(b_i)^2 + 5p(1 - p)^4 L_1(W, b) + O(p^2),$$

where the error term is uniform on bounded sets of (W, b) . Since every term of the loss is nonnegative,

$$L(W, b) \geq (1 - p)^5 \sum_{i=1}^5 \text{ReLU}(b_i)^2,$$

while $L(W, 0) = O(p)$ for any fixed W . So minimizers satisfy $\text{ReLU}(b_i)^2 = O(p)$. Heuristically, the limit problem as $p \rightarrow 0$ is Problem 11.1 with the constraint $b \leq 0$. Numerical minimization of L_1 finds a regular pentagon with norms $\|w_i\| \approx 1.1705$ and biases $b_i \approx -0.2823$, so the constraint $b \leq 0$ is inactive. Heuristically, if Problem 11.1 holds with a Hessian that is nondegenerate in the directions transverse to the G -orbit, then an implicit function argument should give part (1) of Problem 11.3, and Problem 11.2 with $J = (0, p_0)$. This has not been carried out.

The next problem asks for the exact price of legibility. Let $N > m \geq 2$. For W with at least two nonzero columns, let $\mu(W)$ be the coherence of its nonzero columns, normalized, and let $\mu(W) = 0$ otherwise. For $c \in [0, 1]$, let

$$P_{N,m,S}(c) = \inf \{L(W, b) : \mu(W) \leq c\}.$$

- Problem 11.4** (MAIS-O4). (1) *Determine $P_{3,2,S}(c)$ and $P_{5,2,S}(c)$ for all $S \in (0, 1)$ and $c \in [0, 1]$.*
 (2) *Prove or disprove that for fixed (N, m, S) there are finitely many points $0 = c_0 < c_1 < \dots < c_r = 1$ such that on each (c_{j-1}, c_j) , $P_{N,m,S}$ agrees with a function which is real-analytic on a neighborhood of $[c_{j-1}, c_j]$.*
 (3) *Determine $c^*(N, m, S) = \inf \{c : P_{N,m,S}(c) = P_{N,m,S}(1)\}$.*

Two anchors are proved. By Proposition 11.8,

$$P_{N,m,S}(0) = (N - m)v(S).$$

Suppose that $c < m^{-\frac{1}{2}}$ and that W has $n > m$ nonzero columns with $\mu(W) \leq c$. Then Theorem 11.3 gives

$$c^2 \geq \frac{n - m}{m(n - 1)}, \quad \text{i.e.,} \quad n \leq N_*(c) = \frac{m(1 - c^2)}{1 - mc^2}.$$

Since each zero column costs at least $v(S)$,

$$P_{N,m,S}(c) \geq (N - N_*(c))v(S) \quad \text{for } c < m^{-\frac{1}{2}}.$$

For $m = 2$, the bound is sharper: n lines in the plane contain two lines at an angle at most $\frac{\pi}{n}$, so $\mu(W) \geq \cos \frac{\pi}{n}$.

Remark (Semialgebraic structure). A subset of $\mathbb{R}^d$ is semialgebraic if it is a finite union of sets defined by finitely many polynomial equations and inequalities, and a function is semialgebraic if its graph is. For fixed S , the loss $L(W, b)$ is a sum, over the supports of x , of integrals of piecewise polynomials over polytopes in $[0, 1]^N$ whose faces depend polynomially on (W, b) . So it is a rational function on each piece of a semialgebraic partition of the parameter space. Hence L is semialgebraic, and so is the set of (W, c) with $\mu(W) \leq c$. The graph of $P_{N,m,S}$ is defined by the formula “ $y \leq L(W, b)$ for all (W, b) with $\mu(W) \leq c$, and for every $\eta > 0$ there is such a (W, b) with $L(W, b) < y + \eta$ ”. By the elimination of quantifiers over $\mathbb{R}$, $P_{N,m,S}$ is semialgebraic; see Bochnak-Coste-Royⁱ. A semialgebraic function of one variable is real-analytic outside finitely many points. Let c_j be one of these points, or 0 or 1. On each side of c_j in $[0, 1]$, the function $P_{N,m,S}$ has a convergent Puiseux expansion

$$P_{N,m,S}(c) = \sum_{k \geq 0} \alpha_k |c - c_j|^{\frac{k}{q}} \quad \text{for some } q \in \mathbb{N}.$$

So the weak form of part (2) of Problem 11.4, with functions which are real-analytic on the open intervals, always holds. The question is whether only integer powers of $c - c_j$ occur. Fractional powers do appear in simple constrained problems: the minimum of $(x - 1)^2$ subject to $x^2 \leq c$ equals $(1 - \sqrt{c})^2$ for $0 \leq c \leq 1$.

The smallest case is a student problem. For $N = 2$, $m = 1$, importances $(1, I)$, and the penalty $R = 2w_1^2 w_2^2$, the problem MAIS-O51 asks for the exact phase diagram of the global minimizers of $L + \lambda R$ in the parameters $(I, S, \lambda) \in (0, \infty) \times (0, 1) \times (0, \infty)$. The candidates are the three phases found by Elhage et al. for $\lambda = 0$: drop the second feature, dedicate the dimension to it, or store the pair antipodally. Everything reduces to one-dimensional integrals of piecewise polynomials.

ⁱJ. Bochnak, M. Coste, and M.-F. Roy, *Real Algebraic Geometry*. Springer, 1998.

CHAPTER 12

Fourier analysis of learned algorithms

Networks trained on algebraic tasks learn algorithms that can be read from their weights, which makes these tasks the test cases of interpretability. The best understood example is addition modulo p , and the algorithm that networks learn for it is Fourier analysis on the cyclic group $\mathbb{Z}/p\mathbb{Z}$. In this chapter, we review Fourier analysis on $\mathbb{Z}/p\mathbb{Z}$, construct exact fits of the quadratic network of Chapter 2 by single-frequency neurons, and describe what is observed in trained networks. We then state the problems MAIS-O62 on the minimal width, MAIS-O63 on the learning coefficients, MAIS-O6 on whether a small learning coefficient forces the Fourier mechanism, and MAIS-O59 on two neurons learning addition modulo 5. We end with a remark on finite groups.

Throughout this chapter, $p \geq 3$ is an odd integer, $\delta_{a+b=c}$ is 1 if $a + b = c$ in $\mathbb{Z}/p\mathbb{Z}$ and 0 otherwise, and $e_c \in \mathbb{R}^p$, $c \in \mathbb{Z}/p\mathbb{Z}$, are the coordinate vectors.

12.1. Fourier analysis on $\mathbb{Z}/p\mathbb{Z}$

The characters of $\mathbb{Z}/p\mathbb{Z}$ are the homomorphisms $\chi_k : \mathbb{Z}/p\mathbb{Z} \rightarrow \{z \in \mathbb{C} : |z| = 1\}$,

$$\chi_k(a) = e^{\frac{2\pi i k a}{p}} \quad \text{for } k \in \mathbb{Z}/p\mathbb{Z}.$$

They are orthogonal, i.e.,

$$\sum_a \chi_k(a) \overline{\chi_l(a)} = p \delta_{k=l}.$$

The discrete Fourier transform of $f : \mathbb{Z}/p\mathbb{Z} \rightarrow \mathbb{C}$ and its inverse are

$$\hat{f}(k) = \sum_{a \in \mathbb{Z}/p\mathbb{Z}} f(a) e^{-\frac{2\pi i k a}{p}} \quad \text{and} \quad f(a) = \frac{1}{p} \sum_{k \in \mathbb{Z}/p\mathbb{Z}} \hat{f}(k) e^{\frac{2\pi i k a}{p}}.$$

The Plancherel identity reads

$$\sum_a |f(a)|^2 = \frac{1}{p} \sum_k |\hat{f}(k)|^2,$$

and the convolution

$$(f * g)(c) = \sum_{a \in \mathbb{Z}/p\mathbb{Z}} f(a) g(c - a)$$

satisfies $\widehat{f * g} = \hat{f} \hat{g}$; see Terrasⁱ. For real functions, it is convenient to group k with $-k$. For $0 \leq k \leq \frac{p-1}{2}$, let

$$\mathcal{F}_k = \text{span} \left\{ \cos \left(\frac{2\pi k \cdot}{p} \right), \sin \left(\frac{2\pi k \cdot}{p} \right) \right\} \subset \mathbb{R}^{\mathbb{Z}/p\mathbb{Z}}.$$

Then $\mathcal{F}_0$ is the space of constants, $\dim \mathcal{F}_k = 2$ for $k \geq 1$, and $\mathbb{R}^{\mathbb{Z}/p\mathbb{Z}}$ is the orthogonal direct sum of $\mathcal{F}_0, \dots, \mathcal{F}_{\frac{p-1}{2}}$. We call k a frequency. A translation $f \mapsto f(\cdot - t)$ acts on $\mathcal{F}_k$ by the rotation through the angle $\frac{2\pi k t}{p}$.

ⁱA. Terras, *Fourier Analysis on Finite Groups and Applications*. Cambridge University Press, 1999.

Modular addition is diagonalized by the Fourier transform. By orthogonality of the characters,

$$\delta_{a+b=c} = \frac{1}{p} \sum_{k \in \mathbb{Z}/p\mathbb{Z}} e^{\frac{2\pi i k(a+b-c)}{p}} = \frac{1}{p} + \frac{2}{p} \sum_{k=1}^{\frac{p-1}{2}} \cos\left(\frac{2\pi k(a+b-c)}{p}\right). \quad (12.1)$$

In other words, the tensor $T(a, b, c) = \delta_{a+b=c}$ is the sum of the p rank-one tensors $\frac{1}{p}\chi_k \otimes \chi_k \otimes \overline{\chi_k}$. It is the kernel of the convolution, i.e.,

$$(f * g)(c) = \sum_{a, b \in \mathbb{Z}/p\mathbb{Z}} T(a, b, c) f(a) g(b).$$

Each frequency can be handled separately, and on each frequency, adding residues amounts to adding angles. With $\alpha = \frac{2\pi ka}{p}$ and $\beta = \frac{2\pi kb}{p}$,

$$\cos(\alpha + \beta) = \cos \alpha \cos \beta - \sin \alpha \sin \beta \quad \text{and} \quad \sin(\alpha + \beta) = \sin \alpha \cos \beta + \cos \alpha \sin \beta.$$

We shall see in Section 12.2 that trained networks find this algorithm.

12.2. Exact fits by single-frequency neurons

Consider the quadratic network of Chapter 2,

$$f_w(a, b) = \sum_{j=1}^H V_j \left(u'_j(a) + u''_j(b) \right)^2 \in \mathbb{R}^p \quad \text{for } (a, b) \in (\mathbb{Z}/p\mathbb{Z})^2,$$

with the lookup tables $u'_j, u''_j \in \mathbb{R}^{\mathbb{Z}/p\mathbb{Z}}$ and the output vectors $V_j \in \mathbb{R}^p$, so that $w \in \mathbb{R}^d$ with $d = 3pH$. Write $V_j(c)$ for the c -th coordinate of V_j , and

$$n_j(a, b) = \left(u'_j(a) + u''_j(b) \right)^2$$

for the output of the j -th neuron. The network is trained with the square loss averaged over all p^2 inputs,

$$L(w) = \frac{1}{2p^2} \sum_{a, b \in \mathbb{Z}/p\mathbb{Z}} |f_w(a, b) - e_{a+b}|^2,$$

and its set of exact fits is $W_0(p, H) = L^{-1}(0)$. The loss L is a polynomial of degree 6 in $3pH$ variables. Let $F_w(a, b, c) = f_w(a, b)(c)$. Since

$$\hat{T}(k_1, k_2, k_3) = p^2 \delta_{k_1=k_2=-k_3},$$

the Plancherel identity on $(\mathbb{Z}/p\mathbb{Z})^3$ gives

$$L(w) = \frac{1}{2p^5} \left(\sum_{k \in \mathbb{Z}/p\mathbb{Z}} \left| \hat{F}_w(k, k, -k) - p^2 \right|^2 + \sum_{(k_1, k_2, k_3) \neq (k, k, -k)} \left| \hat{F}_w(k_1, k_2, k_3) \right|^2 \right).$$

So an exact fit must produce the diagonal Fourier coefficients and cancel all others.

A neuron j is called active if $V_j n_j$ is not identically zero, and single-frequency if there is a frequency k such that $u'_j, u''_j \in \mathcal{F}_0 + \mathcal{F}_k$. Constants must be allowed, since $(u'_j, u''_j) \mapsto (u'_j + s, u''_j - s)$ does not change n_j .

Proposition 12.1. *If $H \geq 2p - 1$, then $W_0(p, H) \neq \emptyset$. Explicitly, for $H = 2p - 1$, index the neurons by 0 and by (k, i) with $1 \leq k \leq \frac{p-1}{2}$ and $1 \leq i \leq 4$. With $\alpha = \frac{2\pi ka}{p}$, $\beta = \frac{2\pi kb}{p}$, and $\gamma = \frac{2\pi kc}{p}$, let*

$$u'_0(a) = u''_0(b) = \frac{1}{2}, \quad V_0(c) = \frac{1}{p} - \frac{2}{p} \sum_{k=1}^{\frac{p-1}{2}} (\cos \gamma + \sin \gamma),$$

$$\begin{aligned}
(u'_{k,1}(a), u''_{k,1}(b)) &= (\cos \alpha, \cos \beta), & (u'_{k,2}(a), u''_{k,2}(b)) &= (\sin \alpha, -\sin \beta), \\
(u'_{k,3}(a), u''_{k,3}(b)) &= (\sin \alpha, \cos \beta), & (u'_{k,4}(a), u''_{k,4}(b)) &= (\cos \alpha, \sin \beta), \\
V_{k,1}(c) = V_{k,2}(c) &= \frac{\cos \gamma}{p}, & V_{k,3}(c) = V_{k,4}(c) &= \frac{\sin \gamma}{p}.
\end{aligned}$$

Then this parameter w^F satisfies

$$f_{w^F}(a, b) = e_{a+b} \quad \text{for all } (a, b) \in (\mathbb{Z}/p\mathbb{Z})^2,$$

and every neuron of w^F is single-frequency. For $H > 2p - 1$, add neurons with $u'_j = u''_j = 0$ and $V_j = 0$.

PROOF. Expanding the squares, $n_0 = 1$ and

$$\begin{aligned}
n_{k,1} + n_{k,2} &= (\cos \alpha + \cos \beta)^2 + (\sin \alpha - \sin \beta)^2 = 2 + 2 \cos(\alpha + \beta), \\
n_{k,3} + n_{k,4} &= (\sin \alpha + \cos \beta)^2 + (\cos \alpha + \sin \beta)^2 = 2 + 2 \sin(\alpha + \beta).
\end{aligned}$$

Hence, by the choice of V_0 ,

$$f_{w^F}(a, b)(c) = \frac{1}{p} + \frac{2}{p} \sum_{k=1}^{\frac{p-1}{2}} (\cos \gamma \cos(\alpha + \beta) + \sin \gamma \sin(\alpha + \beta)) = \frac{1}{p} + \frac{2}{p} \sum_{k=1}^{\frac{p-1}{2}} \cos(\alpha + \beta - \gamma),$$

which is $\delta_{a+b=c}$ by (12.1). $\square$

The squares $\cos^2 \alpha$, $\sin^2 \alpha$, $\cos^2 \beta$, $\sin^2 \beta$ produced by single neurons carry the frequency $2k$, which is not part of the answer. The neurons come in pairs with a common output vector so that these terms cancel by $\cos^2 + \sin^2 = 1$. This is why four neurons per frequency are used.

The fit w^F is not isolated. Each neuron has the scaling and the translation

$$(u'_j, u''_j, V_j) \mapsto (tu'_j, tu''_j, t^{-2}V_j) \quad \text{for } t > 0 \quad \text{and} \quad (u'_j, u''_j) \mapsto (u'_j + s, u''_j - s),$$

which do not change $V_j n_j$. In particular, $W_0(p, H)$ is unbounded. Moreover, replacing (α, β) by $(\alpha - \phi, \beta + \phi)$ in the neurons $(k, 1)$ and $(k, 2)$, or in the neurons $(k, 3)$ and $(k, 4)$, does not change the sums $n_{k,1} + n_{k,2}$ and $n_{k,3} + n_{k,4}$, since they depend only on $\alpha + \beta$. These rotations extend the symmetry $(a, b) \mapsto (a - t, b + t)$ of the addition table to real t on the frequency k .

Trained networks find this mechanism. Nanda et al. [NCL] reverse-engineered a one-layer transformer trained on addition modulo $p = 113$. The network embeds a and b by $\cos(\frac{2\pi ka}{p})$ and $\sin(\frac{2\pi ka}{p})$ for a few key frequencies k , computes $\cos(\frac{2\pi k(a+b)}{p})$ and $\sin(\frac{2\pi k(a+b)}{p})$ by the addition formulas, and outputs logits approximately of the form

$$c \mapsto \sum_{k \in F} \kappa_k \cos\left(\frac{2\pi k(a+b-c)}{p}\right) \quad \text{with } \kappa_k > 0,$$

where $F = \{14, 35, 41, 42, 52\}$ is the set of the five key frequencies, out of the 56 available. The logits are maximal exactly at $c = a + b$; see Figure 12.1. Ablations confirm this reading: removing the key frequencies destroys the performance of the network, while removing all other frequencies does not hurt it. A different random initialization gives a different set F . Morwani et al.ⁱ proved that, for a quadratic network of the above kind and of sufficient width, in every network that maximizes the margin of classification (Section 1.2), normalized by a norm adapted to the architecture, every active neuron is single-frequency and every frequency $1 \leq k \leq \frac{p-1}{2}$ is

ⁱD. Morwani, B. Edelman, C.-A. Oncescu, et al., *Feature emergence via margin maximization: case studies in algebraic tasks*. International Conference on Learning Representations, 2024.

used. So the mechanism is forced by one selection principle, but the sparse set of frequencies observed in practice is not.

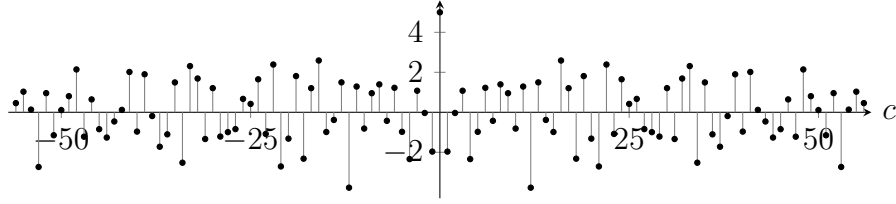


FIGURE 12.1. The logits $c \mapsto \sum_{k \in F} \cos(\frac{2\pi kc}{113})$ on $\mathbb{Z}/113\mathbb{Z}$ for $a + b = 0$ and $\kappa_k = 1$, with $F = \{14, 35, 41, 42, 52\}$: five frequencies give the maximum 5 at $c = 0$ only, while the next largest value is about 2.59

12.3. The minimal width

Proposition 12.1 gives exact fits of width $2p - 1$. How many neurons are necessary?

Problem 12.1 (MAIS-O62). *Let $H_{\min}(p)$ be the smallest H such that $W_0(p, H) \neq \emptyset$. Determine $H_{\min}(p)$.*

This is a constrained tensor rank problem. A complex multiplication can be done with three real multiplications instead of four: for $z = x + iy$ and $z' = x' + iy'$,

$$\operatorname{Re}(zz') = xx' - yy' \quad \text{and} \quad \operatorname{Im}(zz') = (x + y)(x' + y') - xx' - yy',$$

a trick that goes back to Gauss. By the Fourier transform, the convolution of real functions with mean zero amounts to $\frac{p-1}{2}$ products of complex numbers. So it can be computed with $\frac{3(p-1)}{2}$ real products. Winogradⁱ proved that this is optimal: if P is a real polynomial of degree n with t distinct irreducible factors, then every bilinear algorithm for the multiplication in $\mathbb{R}[x]/(P)$, i.e., every identity

$$gh = \sum_{j=1}^r \ell_j(g)m_j(h)z_j \quad \text{for all } g, h \in \mathbb{R}[x]/(P)$$

with linear forms ℓ_j, m_j and elements z_j , has

$$r \geq 2n - t.$$

In the language of tensors, such an identity decomposes the tensor of the multiplication into the r rank-one tensors $\ell_j \otimes m_j \otimes z_j$, so the smallest r is the tensor rank. The case of Winograd's theorem used below has a short proof by substitution, which shows how the field $\mathbb{R}$ enters.

Lemma 12.2. *Let $t \geq 1$, and consider $\mathbb{C}^t$ with the componentwise product as an algebra of dimension $2t$ over $\mathbb{R}$. Suppose that there are real linear forms $\ell_1, \dots, \ell_r$ and $m_1, \dots, m_r$ on $\mathbb{C}^t$ and $z_1, \dots, z_r \in \mathbb{C}^t$ such that*

$$gh = \sum_{j=1}^r \ell_j(g)m_j(h)z_j \quad \text{for all } g, h \in \mathbb{C}^t.$$

Then $r \geq 3t$.

ⁱS. Winograd, *Some bilinear forms whose multiplicative complexity depends on the field of constants*. Math. Systems Theory 10 (1977), 169–180.

PROOF. Let $\pi_i : \mathbb{C}^t \rightarrow \mathbb{C}$ be the i -th coordinate, and let E_i be the plane of the real linear forms $\operatorname{Re}(\alpha\pi_i)$ with $\alpha \in \mathbb{C}$. Then the dual space of $\mathbb{C}^t$ is $E_1 \oplus \cdots \oplus E_t$. The forms ℓ_j span it, since otherwise some $g \neq 0$ satisfies $\ell_j(g) = 0$ for all j , and then $g = g \cdot (1, \dots, 1) = 0$.

We first choose t linearly independent forms among the ℓ_j whose span Λ contains no E_i . Suppose that Λ_s is spanned by $s < t$ of them and contains no E_i , and let I be the set of i with $\Lambda_s \cap E_i \neq \{0\}$. Non-zero vectors in different E_i are linearly independent, so $|I| \leq s$. Each E_i with $i \in I$ meets Λ_s , so it adds at most one dimension to Λ_s . Hence the space

$$Y = \Lambda_s + \sum_{i \in I} E_i$$

has dimension at most $s + |I| \leq 2s < 2t$. Choose $\ell_j \notin Y$, and suppose that $\Lambda_s + \mathbb{R}\ell_j$ contains some E_i . If $i \notin I$, this is impossible, since Λ_s has codimension one in $\Lambda_s + \mathbb{R}\ell_j$, so it meets the plane E_i . If $i \in I$, then $\Lambda_s + E_i$ has dimension $s + 1$, so it equals $\Lambda_s + \mathbb{R}\ell_j$, and $\ell_j \in Y$, which is excluded. So $\Lambda_{s+1} = \Lambda_s + \mathbb{R}\ell_j$ contains no E_i , and we continue until $s = t$.

Let U be the common kernel of the t chosen forms. Then $\dim U = t$, and the forms that vanish on U are exactly those in Λ . If $U \subset \ker \pi_i$, then $E_i \subset \Lambda$, which is excluded. So each $U \cap \ker \pi_i$ is a proper subspace of U . A real vector space is not a finite union of proper subspaces, so there is $g_0 \in U$ with $\pi_i(g_0) \neq 0$ for all i , i.e., g_0 is invertible. Then $h \mapsto g_0 h$ is invertible, of rank $2t$. It equals $h \mapsto \sum_j \ell_j(g_0) m_j(h) z_j$, in which the t chosen terms vanish. So it is a sum of $r - t$ maps of rank at most one, and $r - t \geq 2t$. $\square$

The map $g \mapsto (\hat{g}(k))_{1 \leq k \leq \frac{p-1}{2}}$ is a real linear isomorphism from the real functions with mean zero onto $\mathbb{C}^{\frac{p-1}{2}}$, since $\hat{g}(0) = 0$ and $\hat{g}(-k) = \overline{\hat{g}(k)}$, and it turns the convolution into the componentwise product. The rank of the single map $h \mapsto g_0 h$ gives only $r \geq 2t$. The substitution $\ell_j(g) = 0$ for t of the forms removes t more products, and it uses that the E_i are planes. Over $\mathbb{C}$, the convolution of complex functions with mean zero is the componentwise product on $\mathbb{C}^{p-1}$, the planes E_i become the lines $\mathbb{C}\pi_i$, and a line may lie in $\Lambda_s + \mathbb{C}\ell_j$ without meeting Λ_s . So the first step breaks down, and only the bound $r \geq p - 1$ remains, which the Fourier transform attains.

In the network, each neuron provides one product $u'_j(a)u''_j(b)$, but also the squares $u'_j(a)^2$ and $u''_j(b)^2$, which have to cancel. The following bounds, and the numerical fits for $p = 5$ mentioned below, are due to Shi and the AI model Claude Opus¹.

Proposition 12.3. *For every odd $p \geq 3$,*

$$\frac{3p-1}{2} \leq H_{\min}(p) \leq 2p-1.$$

Moreover, $H_{\min}(3) = 4$.

PROOF. The upper bound is Proposition 12.1. For the lower bound, let $w \in W_0(p, H)$, so that

$$\delta_{a+b=c} = \sum_{j=1}^H V_j(c) n_j(a, b).$$

For a function $M(a, b)$, let ΠM be M minus its means in a and in b plus its total mean. Then Π annihilates the functions of a alone and of b alone. Writing $\tilde{u}$ for u minus its mean, we have

$$\Pi n_j = 2\tilde{u}'_j \otimes \tilde{u}''_j \quad \text{and} \quad \Pi \delta_{a+b=c} = \delta_{a+b=c} - \frac{1}{p}.$$

¹I. Shi and Claude Opus, *A width bound for quadratic network modular addition*. Progress report on MAIS-O62, 2026.

Hence

$$\delta_{a+b=c} - \frac{1}{p} = 2 \sum_{j=1}^H V_j(c) \tilde{u}'_j(a) \tilde{u}''_j(b). \quad (12.2)$$

Let A be the space of real functions on $\mathbb{Z}/p\mathbb{Z}$ with mean zero. For $g, h \in A$, multiply (12.2) by $g(a)h(b)$, sum over (a, b) , and subtract the mean in c . We obtain

$$g * h = 2 \sum_{j=1}^H \langle \tilde{u}'_j, g \rangle \langle \tilde{u}''_j, h \rangle \tilde{V}_j,$$

a bilinear algorithm with H products for the convolution on A . The map $g \mapsto \sum_a g(a)x^a$ identifies $(\mathbb{R}^{\mathbb{Z}/p\mathbb{Z}}, *)$ with $\mathbb{R}[x]/(x^p - 1)$, and, followed by the reduction modulo P , identifies A with $\mathbb{R}[x]/(P)$, where

$$P(x) = 1 + x + \cdots + x^{p-1} = \prod_{k=1}^{\frac{p-1}{2}} \left(x^2 - 2 \cos \frac{2\pi k}{p} x + 1 \right)$$

has $\frac{p-1}{2}$ distinct irreducible factors. By Winograd's theorem, which in this case is Lemma 12.2 with $t = \frac{p-1}{2}$, the number of neurons j with $\tilde{u}'_j \neq 0$ and $\tilde{u}''_j \neq 0$ is at least

$$2(p-1) - \frac{p-1}{2} = \frac{3(p-1)}{2}.$$

Suppose that $H = \frac{3(p-1)}{2}$. Then $\tilde{u}'_j \neq 0$ and $\tilde{u}''_j \neq 0$ for all j , and the matrices $\tilde{u}'_j \otimes \tilde{u}''_j$ are linearly independent. Otherwise, one of them could be removed from (12.2), which would give an algorithm with fewer products. Since $\sum_c \delta_{a+b=c} = 1$, we have

$$\sum_{j=1}^H s_j n_j = 1 \quad \text{with } s_j = \sum_{c \in \mathbb{Z}/p\mathbb{Z}} V_j(c).$$

Applying Π gives $\sum_j s_j \tilde{u}'_j \otimes \tilde{u}''_j = 0$, so all $s_j = 0$, and then $1 = 0$, a contradiction. Hence

$$H \geq \frac{3(p-1)}{2} + 1 = \frac{3p-1}{2}.$$

For $p = 3$, the lower bound is 4. Let $\delta_m(a) = \delta_{a=m}$. Consider the three neurons

$$u'_m = u''_m = \delta_m - \frac{5}{12} \quad \text{and} \quad V_m(c) = \frac{3\delta_{c=2m} - 1}{2} \quad \text{for } m \in \mathbb{Z}/3\mathbb{Z},$$

and a fourth neuron with $u' = u'' = \frac{1}{2}$ and $V(c) = \frac{1}{3}$ for all c . Then

$$n_m = 2\delta_m(a)\delta_m(b) - \frac{2}{3}(\delta_m(a) + \delta_m(b)) + \frac{25}{36} \quad \text{and} \quad \sum_{m \in \mathbb{Z}/3\mathbb{Z}} V_m(c) = 0.$$

Writing $h = 2c$, so that $c = 2h$ in $\mathbb{Z}/3\mathbb{Z}$, we find

$$f_w(a, b)(c) = 1 + 3\delta_{a=b=h} - \delta_{a=h} - \delta_{b=h} - \delta_{a=b}.$$

If $a = b = h$, this is 1; if exactly one of $a = b$, $a = h$, $b = h$ holds, it is 0; and if a, b, h are distinct, it is 1. In the last case, $a + b + h = 0 + 1 + 2 = 0$, so $a + b = -h = 2h = c$. In all cases, $f_w(a, b)(c) = \delta_{a+b=c}$. $\square$

For $p = 3$, every function on $\mathbb{Z}/3\mathbb{Z}$ lies in $\mathcal{F}_0 + \mathcal{F}_1$, so the neurons above are single-frequency as well, although they look like lookup tables of single inputs. For $p \geq 5$, the problem is open. No exact fit of width less than $2p - 1$ is known. Approximate fits of width 8 for $p = 5$ have been found numerically, with residuals at the level of rounding errors, but this does not prove $H_{\min}(5) \leq 8$. A concrete project is to decide whether $W_0(5, 7)$ and $W_0(5, 8)$ are empty, e.g., by

Gröbner bases or by the certificates of the real Nullstellensatz in the remark below. The difficulty is the one seen in Proposition 12.1: each neuron also produces the squares $u'_j(a)^2$ and $u''_j(b)^2$, which carry the frequency $2k$ for a single-frequency neuron, and these must cancel inside the architecture.

Remark (Approximate fits and border rank). Small residuals do not imply exact fits, since the set of the outputs of networks of width H need not be closed. For instance, let u', u'', g', g'' be functions on $\mathbb{Z}/p\mathbb{Z}$, let $V, W \in \mathbb{R}^p$, and let $s > 0$. Write $x(a, b) = u'(a) + u''(b)$ and $y(a, b) = g'(a) + g''(b)$. The two neurons $(u' + sg', u'' + sg'', s^{-1}V + W)$ and $(u', u'', -s^{-1}V)$ together compute

$$(s^{-1}V + W)(x + sy)^2 - s^{-1}Vx^2 = 2Vxy + Wx^2 + s(Vy^2 + 2Wxy + sWy^2),$$

which converges as $s \rightarrow 0$, while the weights blow up. Suppose that V, W are linearly independent, and so are $\tilde{u}', \tilde{g}'$ and $\tilde{u}'', \tilde{g}''$, in the notation of the proof of Proposition 12.3. Then the limit is not the output of two neurons. Indeed, for $\eta \in \mathbb{R}^p$ and a network of width H , the matrix $\sum_c \eta(c) \Pi f_w(\cdot, \cdot)(c)$ lies in the span of the H matrices $\tilde{u}'_j \otimes \tilde{u}''_j$ of rank at most one. For the limit, it is

$$2\langle \eta, V \rangle (\tilde{u}' \otimes \tilde{g}'' + \tilde{g}' \otimes \tilde{u}'') + 2\langle \eta, W \rangle \tilde{u}' \otimes \tilde{u}'',$$

i.e., in the bases $\tilde{u}', \tilde{g}'$ and $\tilde{u}'', \tilde{g}''$, these matrices form the plane

$$\left\{ \begin{pmatrix} \alpha & \beta \\ \beta & 0 \end{pmatrix} : \alpha, \beta \in \mathbb{R} \right\}.$$

If the limit were the output of two neurons, this plane would lie in the span of two matrices of rank at most one, so it would be spanned by two matrices of rank one. But only the matrices with $\beta = 0$ in the plane have rank one. This is the phenomenon of border rank in algebraic complexity. So approximate fits with residuals tending to zero prove $H_{\min}(p) \leq H$ if their weights stay bounded. Indeed, a limit point then lies in $W_0(p, H)$, since L is continuous.

The field matters for computer algebra as well. A Gröbner basis decides whether the equations $f_w(a, b) = e_{a+b}$ have a complex solution $w \in \mathbb{C}^{3pH}$. By Hilbert's Nullstellensatz, there is none if and only if 1 lies in the ideal generated by the p^3 polynomials $f_w(a, b)(c) - \delta_{a+b=c}$. Over $\mathbb{C}$, only $p - 1$ products are needed, so the argument of Proposition 12.3 gives only $H \geq p$. So complex solutions may exist although $W_0(p, H) = \emptyset$. By the real Nullstellensatz, $W_0(p, H) = \emptyset$ if and only if there are real polynomials q_i and r_{abc} in w such that

$$1 + \sum_i q_i(w)^2 = \sum_{a,b,c} r_{abc}(w) (f_w(a, b)(c) - \delta_{a+b=c}).$$

Such certificates of bounded degree can be searched for by semidefinite programming.

12.4. Learning coefficients and the Fourier mechanism

By Chapter 6, Bayesian learning concentrates, as $n \rightarrow \infty$, on the exact fits of smallest local learning coefficient $\lambda(w)$, defined by the volume of $\{L < \varepsilon\}$ near w . If λ can be computed for a task whose learned mechanism is known, it predicts which algorithm the posterior prefers. Modular addition is the natural test case. Since $W_0(p, H)$ is unbounded, we work in a ball. Let $R_p = 1 + |w^F|$, where $|\cdot|$ is the Euclidean norm on $\mathbb{R}^{3p(2p-1)}$. Let $w_H^F \in W_0(p, H)$ be w^F completed by zero neurons. For $R \geq R_p$, let

$$\lambda_{\min}(p, H; R) = \min \{ \lambda(w) : w \in W_0(p, H), |w| \leq R \}.$$

Problem 12.2 (MAIS-O63). *Let $p \geq 3$ be odd and $H \geq 2p - 1$.*

(1) *Determine $\lambda(w_H^F)$ and the multiplicity $m(w_H^F)$.*

(2) For $R \geq R_p$, determine $\lambda_{\min}(p, H; R)$, the multiplicities of the minimizers, and whether the answer depends on R .

The following bounds show what the symmetries give. Consider the map

$$\mathbf{F} : \mathbb{R}^{3pH} \rightarrow \mathbb{R}^{p^3}, \quad w \mapsto (f_w(a, b)(c))_{a,b,c}.$$

Then

$$L(w) = \frac{1}{2p^2} |\mathbf{F}(w) - T|^2.$$

Proposition 12.4. *Let $H = 2p - 1$ and $d = 3p(2p - 1)$, and let r be the rank of the differential of $\mathbf{F}$ at $w^F \in W_0(p, 2p - 1)$. Then*

$$\frac{r}{2} \leq \lambda(w^F) \leq \frac{d - (5p - 3)}{2} = 3p^2 - 4p + \frac{3}{2}.$$

PROOF. For the upper bound, apply first the $p - 1$ rotations, then the $2p - 1$ translations, and then the $2p - 1$ scalings of Section 12.2 to w^F . This gives a smooth map Ψ from a neighborhood of 0 in $\mathbb{R}^{5p-3}$ to $W_0(p, 2p - 1)$, with $\Psi(0) = w^F$. Its differential at 0 is injective. Indeed, the scaling of the neuron j has the tangent vector $(u'_j, u''_j, -2V_j)$ in the block of the neuron j , and every V_j is non-zero, e.g., $V_0(0) = \frac{2-p}{p}$. The translation of the neuron j has the tangent vector $(1, -1, 0)$ in this block. The rotation of the neurons $(k, 1)$ and $(k, 2)$ has the tangent vector with the components $(\sin \alpha, -\sin \beta)$ and $(-\cos \alpha, -\cos \beta)$ in their u -blocks and no V -component, and similarly for $(k, 3)$ and $(k, 4)$. The V -components separate the scalings, and in each block, the translation is constant while the rotation has the frequency k , so the vectors are linearly independent. Hence a neighborhood of w^F in the image of Ψ is a submanifold $\Sigma \subset W_0$ of dimension $5p - 3$. Since $L \geq 0$ is smooth and vanishes on Σ , Taylor's formula gives

$$L(w) \leq C \operatorname{dist}(w, \Sigma)^2 \quad \text{near } w^F.$$

So $\{L < \varepsilon\}$ contains a tube of radius $(\frac{\varepsilon}{C})^{\frac{1}{2}}$ around a piece of Σ , whose volume is at least $c\varepsilon^{\frac{d-(5p-3)}{2}}$, and the volume law gives the upper bound.

For the lower bound, let Q be the orthogonal projection of $\mathbb{R}^{p^3}$ onto the range of $D\mathbf{F}(w^F)$, and let $z(w) = Q(\mathbf{F}(w) - T) \in \mathbb{R}^r$ in an orthonormal basis of the range. Then z is a submersion at w^F , so it can be completed to local coordinates (z, y) near w^F . Since $L \geq \frac{1}{2p^2}|z|^2$, the set $\{L < \varepsilon\}$ lies in $\{|z| < p\sqrt{2\varepsilon}\}$ near w^F , whose volume is at most $C\varepsilon^{\frac{r}{2}}$. Hence $\lambda(w^F) \geq \frac{r}{2}$. $\square$

A numerical computation of r gives $r = 27, 105, 229, 399$ for $p = 3, 5, 7, 9$, i.e.,

$$r = \frac{23p^2 - 28p - 15}{4} \quad \text{for } p = 3, 5, 7, 9.$$

For $p = 3$, $r = 27 = p^3$, so $\mathbf{F}$ is a submersion at w^F , $W_0(3, 5)$ is a smooth submanifold of dimension 18 near w^F , and $\lambda(w^F) = \frac{27}{2}$ with $m = 1$. For $p = 5$, the bounds read $52.5 \leq \lambda(w^F) \leq 56.5$. For these p , the kernel of $D\mathbf{F}(w^F)$ has dimension

$$d - r = \frac{(p+1)(p+15)}{4},$$

which exceeds $5p - 3$. So there are infinitesimal deformations of w^F that do not come from the known symmetries, and Problem 12.2(1) asks, in particular, which of them integrate to curves in W_0 .

The gap between the two bounds is decided on the kernel of $D\mathbf{F}(w^F)$, by a Lyapunov-Schmidt reduction: we solve the equations $\mathbf{F} = T$ in the directions of the range, and we keep the remaining equations as a function on the kernel.

Lemma 12.5. *Let $w_* \in W_0(p, H)$, let $r = \text{rank } D\mathbf{F}(w_*)$ and $K = \ker D\mathbf{F}(w_*)$, and let Q be the orthogonal projection onto the range of $D\mathbf{F}(w_*)$. For $y \in K$ near 0, let $w(y) = w_* + y + \xi(y)$ be the unique point near w_* with $\xi(y) \perp K$ and $Q(\mathbf{F}(w(y)) - T) = 0$, and let*

$$\kappa(y) = (\mathbf{I} - Q)(\mathbf{F}(w(y)) - T).$$

Then

$$\kappa(y) = \frac{1}{2}(\mathbf{I} - Q)D^2\mathbf{F}(w_*)[y, y] + O(|y|^3).$$

If $\kappa \not\equiv 0$, then $\lambda(w_*) = \frac{r}{2} + \lambda_0$ and $m(w_*) = m_0$, where (λ_0, m_0) are the local invariants of $|\kappa|^2$ at 0, as a function on K . If $\kappa \equiv 0$, i.e., if $W_0(p, H)$ contains a submanifold of dimension $d - r$ through w_* , then $\lambda(w_*) = \frac{r}{2}$ and $m(w_*) = 1$.

PROOF. Since $D\mathbf{F}(w_*)$ maps $K^\perp$ isomorphically onto its range, the point $w(y)$ exists by the implicit function theorem, and ξ is real-analytic with $\xi(0) = 0$ and $D\xi(0) = 0$. So $\xi(y) = O(|y|^2)$. Since $D\mathbf{F}(w_*)y = 0$ and $(\mathbf{I} - Q)D\mathbf{F}(w_*) = 0$, Taylor's formula gives the expansion of κ .

Near w_* , we use the local coordinates (z, y) , where $z = Q(\mathbf{F}(w) - T)$ and y is the orthogonal projection of $w - w_*$ onto K . They are local coordinates, since $D\mathbf{F}(w_*)$ is injective on $K^\perp$. Then

$$2p^2L(w) = |z|^2 + |\tilde{\kappa}(z, y)|^2 \quad \text{with } \tilde{\kappa}(z, y) = (\mathbf{I} - Q)(\mathbf{F}(w) - T).$$

Here $\tilde{\kappa}(0, y) = \kappa(y)$, and $D\tilde{\kappa}(0, 0) = 0$, since $(\mathbf{I} - Q)D\mathbf{F}(w_*) = 0$. So

$$|\tilde{\kappa}(z, y) - \kappa(y)| \leq C|z|(|z| + |y|).$$

Hence $2p^2L \leq 3|z|^2 + 2|\kappa(y)|^2$ near w_* , while $|z| < p\sqrt{2\varepsilon}$ and $|\kappa(y)| \leq C\sqrt{\varepsilon}$ on $\{L < \varepsilon\}$. For a small $\delta > 0$, let $V_\kappa(s)$ be the volume of $\{y \in K : |y| < \delta, |\kappa(y)|^2 < s\}$. Then the volume of $\{L < \varepsilon\}$ near w_* lies between

$$c\varepsilon^{\frac{r}{2}}V_\kappa(c\varepsilon) \quad \text{and} \quad C\varepsilon^{\frac{r}{2}}V_\kappa(C\varepsilon).$$

If $\kappa \not\equiv 0$, then V_κ satisfies the volume law with the exponents λ_0 and m_0 . If $\kappa \equiv 0$, then V_κ is constant. In both cases, the exponents of L follow, as in the proof of Lemma 6.6(1).

If $\kappa \equiv 0$, then $y \mapsto w(y)$ parametrizes a submanifold of $W_0(p, H)$ of dimension $d - r$ through w_* . Conversely, $\mathbf{F}$ is constant on such a submanifold, so its tangent space at w_* lies in K , and hence equals K . Near w_* , the submanifold is then a graph $\{w_* + y + \eta(y)\}$ over K with $\eta(y) \perp K$. Since $Q(\mathbf{F} - T) = 0$ on it, $\eta = \xi$, and $\kappa \equiv 0$. $\square$

Differentiating $\mathbf{F} = T$ twice along a curve in $W_0(p, H)$ through w_* with the tangent vector y gives $y \in K$ and

$$(\mathbf{I} - Q)D^2\mathbf{F}(w_*)[y, y] = 0.$$

So the quadratic part of κ is the first obstruction for an infinitesimal deformation $y \in K$ to integrate to a curve of exact fits. For w^F and $p = 5, 7$, a numerical computation shows that this quadratic form vanishes on K up to rounding errors, that is,

$$|(\mathbf{I} - Q)D^2\mathbf{F}(w^F)[y, y]| \leq 10^{-13}|y|^2 \quad \text{for } y \in K,$$

while $|D^2\mathbf{F}(w^F)[y, y]|$ is of the order of $|y|^2$. Moreover, for random $y \in K$ with $|y| \leq 0.1$, Newton's method in its least-squares form (Gauss-Newton), in the unknown $\xi \perp K$, solves $\mathbf{F}(w^F + y + \xi) = T$ in at most three steps, with a residual below 10^{-14} and $|\xi| \leq 0.1|y|^2$, and $D\mathbf{F}$ has rank r at the solution. So $\kappa(y) = 0$ up to rounding errors. Exact fits are found in the same way for random $y \in K$ orthogonal to the tangent vectors of the known symmetries. This suggests that $W_0(p, 2p - 1)$ is a submanifold of dimension $d - r$ near w^F . If so, then Lemma 12.5 gives $\lambda(w^F) = \frac{r}{2}$ and $m(w^F) = 1$, e.g., $\lambda(w^F) = 52.5$ for $p = 5$, and the lower bound of Proposition 12.4 is sharp. By Lemma 12.5, a proof amounts to the construction of a family of exact fits of

dimension $d - r$ through w^F , with $\frac{p^2 - 4p + 27}{4}$ parameters beyond the $5p - 3$ known symmetries, e.g., 8 for $p = 5$ and 12 for $p = 7$.

By the remark on oscillatory integrals in Section 6.3, $\lambda(w^F)$ and $m(w^F)$ are also the exponents in the decay

$$\int e^{i\tau L(w)} \varphi(w) dw \sim C \tau^{-\lambda(w^F)} (\log \tau)^{m(w^F)-1} \quad \text{as } \tau \rightarrow \infty,$$

for $\varphi \geq 0$ supported in a small neighborhood of w^F with $\varphi(w^F) > 0$. One may look for coordinates adapted to the frequencies, guided by the Plancherel form of L , in which L becomes non-degenerate with respect to its Newton polyhedron, so that Varchenko's theorem applies. This is a suggestion, not a known method. The geometric question is whether singular learning theory selects the Fourier mechanism.

Question 12.1 (MAIS-O6). *Let $p \geq 3$ be odd and $H \geq 2p - 1$. Is every active neuron of every $w \in W_0(p, H)$ with $|w| \leq R_p$ and $\lambda(w) = \lambda_{\min}(p, H; R_p)$ single-frequency?*

A positive answer would show that geometric simplicity, i.e., a small learning coefficient, forces a mechanism that can be read from the weights. A negative answer would show that the two notions of simplicity part ways. For $p = 3$, the question is vacuous, since $\mathcal{F}_0 + \mathcal{F}_1 = \mathbb{R}^3$. The first non-vacuous case is $p = 5$ and $H = 9$, where $d = 135$. The only mechanism we know that favors the Fourier fits is heuristic: the rotations of Section 12.2 are continuous symmetries that exist because of the frequency structure, and more symmetries mean a larger set of nearby exact fits and so a smaller λ . The theorem of Morwani et al. above shows that single-frequency neurons are forced by margin maximization. Margin and learning coefficient are different optima, so it does not answer Question 12.1.

The question also has a harmonic-analysis side. A single-frequency lookup table is sparse in frequency: its Fourier transform is supported in at most three points. By the uncertainty principles on $\mathbb{Z}/p\mathbb{Z}$ in Chapter 11, it cannot be sparse in space: a non-zero $u \in \mathcal{F}_0 + \mathcal{F}_k$ satisfies $|\text{supp } u| \geq \frac{p}{3}$ by Theorem 11.5, and $|\text{supp } u| \geq p - 2$ by Tao's inequality if p is prime. Directly, for p prime, a non-constant $u \in \mathcal{F}_0 + \mathcal{F}_k$ has the form $u(a) = c_0 + c_1 \cos(\frac{2\pi ka}{p} + \phi)$, and the points $\frac{2\pi ka}{p}$ are distinct modulo 2π , so u takes each value at most twice. So a Fourier neuron responds to essentially every input, whereas a memorizing neuron responds to few inputs. Question 12.1 asks on which side of the uncertainty principle the singular geometry places the network. Problem 12.1 and Question 12.1 are connected: below $H_{\min}(p)$ there is nothing to select from, and the example for $p = 3$ shows that at the minimal width, exact fits need not look like Proposition 12.1.

12.5. Which frequencies are learned

The previous problems concern the Bayesian posterior. For gradient training, the question is which frequencies the neurons choose, and with what probability. The smallest case is addition modulo 5 with two neurons. There are two frequencies, $k = 1$ and $k = 2$, and the network

$$f_w(a, b) = \sum_{i=1}^2 V_i (u'_i(a) + u''_i(b))^2 \in \mathbb{R}^5 \quad \text{with } w = (u'_i, u''_i, V_i)_{i=1,2} \in \mathbb{R}^{30}$$

is trained by the gradient flow $\dot{w}_t = -\nabla \mathcal{L}(w_t)$ of the cross-entropy loss

$$\mathcal{L}(w) = -\frac{1}{25} \sum_{a,b \in \mathbb{Z}/5\mathbb{Z}} \log \frac{e^{f_w(a,b)(a+b)}}{\sum_c e^{f_w(a,b)(c)}},$$

without weight decay, from w_0 with independent standard Gaussian coordinates. Let $\tilde{u}'_i, \tilde{u}''_i, \tilde{V}_i$ be u'_i, u''_i, V_i minus their means, where V_i is regarded as a function on $\mathbb{Z}/5\mathbb{Z}$. Let P_k be the orthogonal projection onto $\mathcal{F}_k$. For $\delta > 0$, the neuron i is called δ -pure of frequency k if

$$|P_k \tilde{u}'_i|^2 + |P_k \tilde{u}''_i|^2 + |P_k \tilde{V}_i|^2 \geq (1 - \delta) \left(|\tilde{u}'_i|^2 + |\tilde{u}''_i|^2 + |\tilde{V}_i|^2 \right).$$

Problem 12.3 (MAIS-O59). (1) Write the gradient field in a real Fourier basis. Classify the stationary directions, i.e., the unit vectors $w \in \mathbb{S}^{29}$ at which $\nabla \mathcal{L}(w)$ is parallel to w , in which each neuron is supported on one frequency after subtracting its mean. Determine their stability transversally to the radial direction, and decide which of the four assignments $(k_1, k_2) \in \{1, 2\}^2$ of frequencies to the two neurons can attract an open set.

- (2) Prove or refute: almost surely, $\frac{w_t}{|w_t|}$ converges as $t \rightarrow \infty$, and there are frequencies k_1, k_2 such that, in the limit, the neuron i is δ -pure of frequency k_i for every $\delta > 0$.
 (3) Compute the probability that $k_1 = k_2$.

Directions are the right objects, since $|w_t|$ may tend to infinity. Weight decay, i.e., adding $\frac{\mu}{2}|w|^2$ with $\mu > 0$ to the loss, would keep $|w_t|$ bounded, since the penalized loss decreases along its gradient flow. Without it, the remark on the cross-entropy in the scores in Section 1.2 applies, with the logits $f_w(a, b)(c)$ and the degree of homogeneity 3. In particular, $|w_t| \rightarrow \infty$ whenever $\mathcal{L}(w_t) \rightarrow 0$. Moreover, by that remark with $n = 25$, if $\mathcal{L}(w) < \frac{\log 2}{25}$, then the margin $\gamma(w)$ is positive, i.e., all 25 inputs are classified correctly. Suppose that $\mathcal{L}(w_{t_0}) < \frac{\log 2}{25}$ for some t_0 . Then, by a theorem of Lyu-Liⁱ on homogeneous models, $|w_t| \rightarrow \infty$, the normalized margin $\gamma(w_t)|w_t|^{-3}$ converges, and every limit point of $\frac{w_t}{|w_t|}$ is the direction of a point which satisfies the first-order conditions, with non-negative Lagrange multipliers, for minimizing $|w|^2$ subject to the constraints $m_{jk}(w) \geq 1$ of that remark. So, in this case, Problem 12.3(2) asks whether $\frac{w_t}{|w_t|}$ has only one limit point, and whether the neurons of this limit are pure. The theorem of Morwani et al. in Section 12.2 concerns the global maximizers of the margin, normalized as there, for wide networks.

Part (1) is a finite computation, suitable for a student. The probability in part (3) determines the law of the pair (k_1, k_2) on the event in part (2). Indeed, the map $a \mapsto 2a$ is an automorphism of $\mathbb{Z}/5\mathbb{Z}$ which exchanges $\mathcal{F}_1$ and $\mathcal{F}_2$, since $2 \cdot 2 = 4 \equiv -1$. Composing all lookup tables and output vectors with it is a permutation of the coordinates of w , which preserves the Gaussian law and $\mathcal{L}$, since $2a + 2b = 2(a + b)$. Hence it commutes with the flow. Together with the exchange of the two neurons, this shows that on the event in part (2), the pairs (1, 1) and (2, 2) have the same probability, and so do (1, 2) and (2, 1).

Remark (Finite groups). For a finite group G , the Fourier analysis of Section 12.1 is replaced by the irreducible representations of G ; see Serreⁱⁱ. Chughtai-Chan-Nanda [CCN] found that networks trained on the multiplication $(a, b) \mapsto ab$ in G learn an algorithm based on a sparse set of irreducible representations, and that which representations appear varies with the random initialization. Interpretability findings transfer across training runs only to the extent that this choice is predictable; Problem 12.3 is the smallest case of this question.

ⁱK. Lyu and J. Li, *Gradient descent maximizes the margin of homogeneous neural networks*. International Conference on Learning Representations, 2020.

ⁱⁱJ.-P. Serre, *Linear Representations of Finite Groups*. Springer, 1977.

CHAPTER 13

Where to start

The problems of Part III are of three kinds. Some are questions of semiclassical analysis and spectral theory, some are questions of harmonic analysis, and some are explicit computations in a few variables, which are the natural first projects for students. In this chapter, we suggest entry points, roughly in the order of their distance from semiclassical analysis, spectral theory, and harmonic analysis. For each of them, we indicate what to read, a first step, and the open problem behind it.

13.1. Metastability of singular wells

The closest problem is MAIS-O74 in Chapter 9. It asks for the Eyring-Kramers law when the starting well is a positive-dimensional singular set and the gate is a Morse saddle. The background is Chapter 8: the formula of Bovier-Eckhoff-Gaynard-Klein for mean hitting times (Section 8.5), and the semiclassical point of view (Section 8.6). Theorem 9.4 proves the law for the starting point distributed according to the equilibrium measure. So the main remaining difficulty is Problem 9.2: an oscillation estimate for the solution of $-\mathcal{L}_\varepsilon w_B = 1$ along a positive-dimensional well. A first step is to compute the constant c_A of Proposition 9.5 for further monomial wells and to compare the resulting prefactor with simulations of the Langevin dynamics. The general problem MAIS-O73, with degenerate gates, is a research program.

The barrier-free companion is MAIS-O75 in Chapter 10. Everything is explicit: the potential is $L(x, y) = \sin^2 x \sin^4 y$ on $\mathbb{T}^2$. Part (a) asks for the time scale of the entropic drift along a stratum, which is predicted by an effective one-dimensional diffusion. Part (b) asks for the spectral gap of the Witten Laplacian P_ε of Lemma 8.4, whose potential vanishes on the whole zero set W_0 , so the harmonic approximation of Section 8.6 does not apply. A first step, suitable for a strong student, is to compute the low-lying spectrum of P_ε numerically in a Fourier basis on $\mathbb{T}^2$, to guess the exponent in part (b), and then to prove the upper bound by test functions.

13.2. Laplace integrals in singular learning theory

The learning coefficient is an exponent in the asymptotics of a Laplace integral, and of the oscillatory integral with the same phase (Section 6.3). Several problems are computations of such asymptotics for explicit phases.

Problem 6.1 asks when a Gaussian localization of the posterior estimates the local learning coefficient. With L in place of L_n , the estimator is $N\mathbb{E}_N[L]$ for a Gibbs measure. So, by Proposition 6.7 and the decomposition after it, it is approximately the average of the learning coefficients of two wells, a line of minima and a degenerate point, weighted by their posterior masses. Hence the problem is a competition between the two wells in a Laplace integral with two large parameters. The heuristic in Chapter 6 predicts the transitional scaling, and a proof needs only Laplace's method with uniform error bounds.

Problem 6.2 asks for the local learning coefficients of reduced rank regression at every point of its optimal set. It needs only linear algebra and elementary volume computations, for which Lemma 6.6 gives the rules, and the case $M = N = H = 2$ is a first project for a student. It is also

the first step of Problem 6.3, and, through part (b) of that problem, of the opposing-staircases conjecture MAIS-O7 in Chapter 7. By Lemma 6.5 and the argument after that conjecture, $\lambda_k \leq 1$ for $k < r$. So, for $M = N = H$, the conjecture concerns the r numbers $\lambda_0, \dots, \lambda_{r-1}$ in $(0, 1]$, while for $M = N = 1$ it fails as stated (Chapter 7).

Problems 7.2 and 7.1 concern Laplace integrals over small windows around the saddle sets of the deep linear network. The rim example of Section 7.3 shows the phenomenon to be understood: near a saddle, the integral is dominated by the boundary of the window in the unstable direction. Parts (a) and (b) of Problem 7.2 are finite-dimensional questions about explicit Laplace integrals.

Question 6.1 on the stratification of the optimal set by the learning coefficient is a question of real-analytic geometry, closer to singularity theory than to analysis. The remark after it indicates that parts (a) and (b) follow from the resolution of singularities. So a first project is to write out this sketch in full, and then to treat the pair (λ, m) , which needs more care.

13.3. Harmonic analysis of superposition

The problems of Chapter 11 are problems on the geometry of nearly orthogonal vectors, and several of them are energy minimization problems on spheres. Problem 11.1 asks whether the regular pentagon minimizes the reconstruction loss of the ReLU toy model with five features in the plane and one active feature. For unit columns and zero biases, it is the minimization of a pair energy on the circle, and the linear programming method of Delsarte-Goethals-Seidel (Section 11.3) reduces it to the construction of a trigonometric polynomial below the potential. Proposition 11.9 gives such a polynomial of degree seven and proves this case. So the first project is the case of free norms and biases in Problem 11.1. The next steps are uniqueness (Problem 11.3) and the full toy model with p in an interval (Problem 11.2). The case $(N, m) = (3, 2)$ of Problem 11.4 and the phase diagram of the model with $N = 2$ and $m = 1$ are explicit computations for students.

Questions 11.2 and 11.3 connect superposition with eigenfunctions. Sparse vectors can be recovered from few random samples of a bounded orthonormal system, and the number of samples is governed by the sup norms of the basis functions. So the recovery of sparse expansions in spherical harmonics, or in random orthonormal bases of the eigenspaces of the Laplacian, depends on the L^∞ and L^p norms of eigenfunctions and on the behavior of random eigenbases. Question 11.1, on the number of random rows of the Fourier matrix of $\mathbb{Z}/N\mathbb{Z}$ that give the restricted isometry property, is a problem of harmonic analysis on cyclic groups.

13.4. Fourier analysis on $\mathbb{Z}/p\mathbb{Z}$

The networks of Chapter 12 learn modular addition through the Fourier transform on $\mathbb{Z}/p\mathbb{Z}$. Problem 12.1 on the minimal width of an exact fit is a problem in the complexity of bilinear maps, and the first open case $p = 5$ is a computation. Lemma 12.2 shows the method behind the lower bound, and the remark on approximate fits and border rank in Section 12.3 says what a computer search has to certify. Part (1) of Problem 12.3 is a finite computation for students. Problem 12.2 and Question 12.1 combine the Fourier transform with the learning coefficient. For w^F and $p = 5, 7$, the numerical evidence after Lemma 12.5 suggests that $\lambda(w^F) = \frac{r}{2}$, and a first step is to construct, for $p = 5$, a family of exact fits through w^F with 8 parameters beyond the known symmetries. The rest of these problems are long-range.

13.5. A reading course

For a one-semester reading course, we suggest Chapters 1–3 for the language, Chapters 4 and 5 for the motivation, Chapters 6 and 8 for the core mathematics, and then one of Chapters 7, 9, 10, 11, and 12, according to the interest of the student.

APPENDIX A

Glossary

We collect the terms from AI and AI safety used in these notes, with the chapters in which they are introduced.

- *Ablation*: removing a component of a trained network to measure its effect on the performance. Chapter 12.
- *Activations*: the values that flow through a neural network on a given input. Chapter 2.
- *Agent*: a system which acts in an environment by a policy. Chapter 2.
- *AI safety*: the study of how to build and deploy AI systems whose behavior is beneficial, and how to prevent the harms of powerful AI. Chapter 4.
- *Bayesian*: the point of view in which probability measures belief, updated by Bayes' rule. Chapter 1.
- *Cross-entropy loss*: the empirical negative log-likelihood loss. Chapter 1.
- *Deep linear network*: a composition of linear layers; with two layers, the map $x \mapsto BAx$. Chapter 3.
- *Developmental interpretability*: the program of reading the formation of internal structure during training from changes in the local learning coefficient. Chapter 5.
- *Empirical loss*: the average loss over the training data. Chapter 1.
- *Entropy*: the average surprise $-\int q \log q$ of a probability distribution. Chapter 1.
- *Feature*: a variable or property which a network represents in its activations. Chapter 4.
- *Fisher information*: the covariance of the score $\nabla_w \log p$; the Hessian of the Kullback-Leibler divergence at the truth. Chapter 1.
- *Free energy*: minus the logarithm of the Bayesian evidence $\int \varphi \prod_j p(X_j | w) dw$. Chapter 6.
- *Generalization*: good predictions on new data; in distribution if the new data come from the training distribution, out of distribution otherwise. Chapter 1.
- *Goal misgeneralization*: a system keeps its capabilities under a distribution shift but pursues a different goal from the intended one. Chapter 4.
- *GPT*: generative pre-trained transformer, a family of language models. Chapter 2.
- *Gradient descent, gradient flow*: the iteration $w_{k+1} = w_k - \eta \nabla L(w_k)$ and its continuous-time limit. Chapter 3.
- *Grokking*: generalization that arrives long after the training loss is small, and abruptly. Chapter 3.
- *Interpretability*: reverse-engineering a trained network into human-understandable structure. Chapter 4.
- *Kullback-Leibler divergence*: the excess $\int q \log \frac{q}{p}$ of the cross-entropy over the entropy. Chapter 1.
- *Langevin dynamics*: the diffusion $dw_t = -\nabla L(w_t) dt + \sqrt{2\varepsilon} dB_t$. Chapters 3 and 8.
- *Language model*: a neural network trained to predict the next token of a text. Chapter 2.
- *Layer, neuron*: a map $x \mapsto \sigma(Ax + b)$, and one of its coordinates. Chapter 2.

- *Learning coefficient*: the exponent λ in the volume law of nearly optimal parameters; locally, the local learning coefficient. Chapter 6.
- *Learning rate*: the step size η of gradient descent. Chapter 3.
- *Legibility*: the extent to which humans can understand how an AI system represents concepts internally. Chapter 4.
- *Likelihood*: the probability that a model with a given parameter assigns to the observed data. Chapter 1.
- *Logits*: the scores which the softmax turns into probabilities. Chapter 1.
- *Loss*: the function that training minimizes. Chapter 1.
- *Margin*: the smallest difference, over the training data, between the logit of the correct output and the other logits. Chapter 1.
- *Maximum likelihood*: the choice of the parameter which maximizes the likelihood. Chapter 1.
- *Minibatch*: the random subset of data used in one step of SGD. Chapter 3.
- *Misalignment*: an AI system pursues a goal different from the intended one. Chapter 4.
- *Neural network*: a composition of layers whose matrices and vectors are the weights. Chapter 2.
- *Noise*: the random part of the outputs, which is not determined by the inputs, e.g., ξ in $y = f_{w_0}(x) + \xi$; also the random part of the steps of training, e.g., the minibatch noise of SGD. Chapters 1 and 3.
- *Overfitting*: fitting the noise in the training data, so that the empirical loss is much smaller than the population loss. Chapter 1.
- *Phase transition*: a moment of training at which the estimated local learning coefficient jumps; in Bayesian learning, a sample size at which the posterior moves to a different region. Chapter 5.
- *Plateau*: a long period of nearly constant loss during training. Chapter 3.
- *Policy*: an agent's decision rule from observations to actions. Chapter 2.
- *Population loss*: the average loss over all possible data, weighted by the data distribution. Chapter 1.
- *Prior, posterior*: the densities on parameter space before and after seeing the data. Chapters 1 and 6.
- *Reinforcement learning*: training an agent by rewarding its actions. Chapter 2.
- *ReLU*: the activation function $z \mapsto \max\{0, z\}$. Chapter 2.
- *Robustness*: good performance on inputs unlike the training and test data. Chapter 4.
- *Sigmoid*: the activation function $z \mapsto \frac{1}{1+e^{-z}}$. Chapter 2.
- *Singular model*: a model which is not identifiable or whose Fisher information is degenerate. Chapter 1.
- *Softmax*: the map from scores to probabilities $f \mapsto \left(\frac{e^{f_y}}{\sum_k e^{f_k}} \right)_y$. Chapter 1.
- *Sparse autoencoder*: a network that reconstructs activations as sparse combinations of learned directions. Chapters 4 and 11.
- *Stochastic gradient descent (SGD)*: gradient descent with minibatch estimates of the gradient. Chapter 3.
- *Superposition*: the representation of more features than dimensions, as nearly orthogonal directions. Chapters 4 and 11.
- *Token*: the atomic unit of text, a word or a piece of a word. Chapter 2.
- *Training*: choosing the parameters of a model from data. Chapters 1 and 3.

- *Transformer*: the dominant architecture of language models, which mixes information across the positions of a text. Chapter 2.
- *Weight decay*: a penalty proportional to $|w|^2$ added to the loss. Chapters 7 and 12.
- *Weights*: the trainable parameters of a neural network. Chapter 2.

Bibliography

- [B] N. Berglund, *Kramers’ law: validity, derivations and generalisations*. Markov Process. Related Fields 19 (2013), 459–490.
- [BEGK] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein, *Metastability in reversible diffusion processes. I. Sharp asymptotics for capacities and exit times*. J. Eur. Math. Soc. (JEMS) 6 (2004), 399–424.
- [CCN] B. Chughtai, L. Chan, and N. Nanda, *A toy model of universality: reverse engineering how networks learn group operations*. Proc. Mach. Learn. Res. 202 (2023), 6243–6267.
- [D] L. Delande, *Sharp spectral gap for some degenerate Witten Laplacians*. Ann. Henri Poincaré (2026).
- [EHO] N. Elhage, T. Hume, C. Olsson, et al., *Toy models of superposition*. arXiv:2209.10652.
- [FR] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.
- [H] X. Han, *The Eyring-Kramers formula and semiclassical tunneling through a Morse gate*. In preparation.
- [HWFR] J. Hoogland, G. Wang, M. Farrugia-Roberts, et al., *Loss landscape degeneracy and stagewise development in transformers*. Trans. Mach. Learn. Res. (2025).
- [L] L. Levine, *Math for AI safety: an invitation for mathematicians*. arXiv:2609.15289.
- [LFW] E. Lau, Z. Furman, G. Wang, et al., *The local learning coefficient: a singularity-aware complexity measure*. Proc. Mach. Learn. Res. 258 (2025), 244–252.
- [NCL] N. Nanda, L. Chan, T. Lieberum, et al., *Progress measures for grokking via mechanistic interpretability*. International Conference on Learning Representations, 2023.
- [W] S. Watanabe, *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.