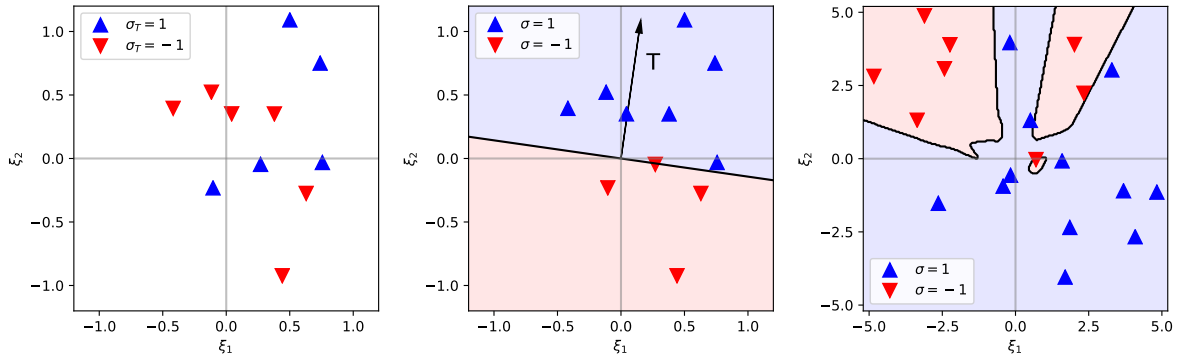
Stochastic Thermodynamics of Learning



Von der Fakultät Mathematik und Physik der Universität Stuttgart
zur Erlangung der Würde eines Doktors der
Naturwissenschaften (Dr. rer. nat.) genehmigte Abhandlung

vorgelegt von

Sebastian Goldt

aus Wiesbaden

Hauptberichter:	Prof. Dr. Udo Seifert
Mitberichter:	Prof. Dr. Christian Holm
Vorsitzende:	Prof. Dr. Stefanie Barz

Tag der Einreichung:	19.12.2017
Tag der mündlichen Prüfung:	09.02.2018

II. Institut für Theoretische Physik der Universität Stuttgart

2018

Contents

Summary

Thermodynamics of information processing

Information processing is ubiquitous in biological systems: Bacteria measure the gradients of the concentration of external nutrients to determine their swimming direction; cells pass faithful copies of their genetic information on to their progeny, and large neural networks are capable of performing complex motor control tasks. These systems are all surprisingly robust, despite the fact that they are operating in noisy environments, and they can be very precise: kinetic proofreading, a mechanism for high-fidelity copying of the genetic code, achieves an error probability of just 3×10^{-8} per letter.

From a thermodynamic perspective, it is the efficiency of these processes which stands out: *E.coli*, a bacterium, is near-perfect in exploiting a given energy budget to adapt its sensory apparatus to changes in its environment, while the human brain consumes less than 20 watts. These facts suggest that thermodynamics played an important role in the evolutionary "design" of the chemical and neural networks that underlie biological computation.

Non-equilibrium statistical physics has made great strides over the last twenty years which have made it fit to study to study fluctuating systems far from equilibrium. Stochastic thermodynamics has emerged from this effort as a mature framework, based on the mathematical theory of Markov processes. It is particularly suited to study the interplay of dissipation and information processing in small systems far from equilibrium. Encouraged by its successful application to the problem of sensing, where cells continuously dissipate free energy to monitor the external state of affairs, here we study a new problem: learning.

Information processing and learning in neural networks

Learning is about extracting models from data and applying them to new problems. In living systems, these processes are implemented in neural networks, where a large number of nerve cells, the neurons, are connected to each other. These neurons communicate with each other using action potentials, the brief electric

pulses that are used universally as the basic token of communication in neural systems. A single neuron can thus be understood as a logical gate within the network: Its input is the activity of all its incoming connections to other neurons, each of which can be firing or not. Its output is also binary: it can either fire another action potential, or stay silent.

Action potentials are transmitted between neurons via synapses. The so-called *weight* of a synapse connecting neurons A and B determines how much neuron B is affected by the inputs it receives from A. The larger the weight, the more likely B is to fire an action potential of its own if it receives an action potential from A. More generally, the set of weights in a network determine how that network processes information. The adaptation of synaptic weights is thought to be the key mechanism for memory formation and learning.

Our approach in this work is to study simplified models of neurons, mostly the famous perceptron, well known from machine learning and statistical physics and inspired by neurobiology. It is the basic building block of neural networks. We focus on the dynamics of the weights, which completely characterise the network, as they evolve to fulfil a certain learning task. Changing these weights in finite time in a noisy environment is a non-equilibrium process and hence incurs a finite thermodynamic cost in the form of dissipated free energy. The key question is this: how much energy must a neural network dissipate in order to learn?

Building a model

We first study a perceptron with N weights which has to build a model for a set of data. Specifically, the network is given a fixed set of P inputs, each of which has a true label which was drawn at random, independently of the input and of the other labels. The goal of learning is to adjust the weights of the network such that the perceptron is able to reconstruct the true labels for as many inputs as possible.

We show that the thermodynamic cost of building such a model, measured by the total entropy production of the weights, is an upper bound on the amount of information that the network can extract from the data, measured by the mutual information between the true labels of the fixed inputs and the labels that the perceptron reconstructed. This is true for perceptrons with any number of weights N learning any number of samples P.

Our bound is sharp, as it can be saturated by a neuron with a single weight learning a single sample, for example. Since the number of connections a neuron makes can be on the order of thousands, we also study learning in the thermodynamic limit, where we let the number of weights and the number of samples both go to infinity while keeping their ratio on the order of one.

Generalising from examples

Perceptrons have the remarkable property that they can learn from examples and generalise what they have learnt to new problems. This feature of perceptrons is most conveniently analysed for a neural network learning a rule. The rules we want to learn are Boolean functions: they take an input and map it to a binary output, the input's true label. The rule to be learned is implemented by another neural network, called the teacher. The perceptron has to infer this rule from a number of examples, *i.e.* from a number of input-label pairs supplied by the teacher, without having direct access to the weights of the teacher network. We are interested in the ability of the network to emulate the function after a training period. How well do the outputs of the student match the correct output of the teacher for any given input?

Generalising from examples is a different problem from building a model for a number of fixed input-label mappings that we discussed in the previous section. The true labels in the first problem were drawn at random for each input, and hence uncorrelated to the inputs and to each other. Hence there is no generalisation error for building a model – if the true label of every input is determined by pure chance, the mappings carry no information about the label of a previously unseen input. Here, the examples are all generated by the same teacher and hence allow the student to generalise from the examples to previously unseen inputs if it can infer the weights of the teacher.

We show that the accuracy with which the neuron is able to apply the rule to previously unseen inputs, given by the mutual information between the true and the predicted label for a randomly chosen input, is constrained by the total entropy production of a single weight during the learning process.

We can refine this bound using concepts from steady state thermodynamics. To that end, we split the total entropy production into two contributions. Applying non-equilibrium constraints such as a fluctuating learning force during online learning leads to *adiabatic* entropy production to maintain the steady state. Driving the system from one steady state to another will additionally lead to a *non-adiabatic* entropy production. This rate goes to zero in the steady state. We show that the non-adiabatic entropy production per weight, which is of order one, is also an upper bound on the mutual information between the true and predicted label.

Our results apply to a wide variety of learning algorithms. For illustration purposes, we analyse the dynamics and thermodynamics of three learning algorithms in particular: Hebbian learning, which was inspired by the neurobiology of memory formation; the celebrated Perceptron algorithm, whose discovery led to a surge in interest in neural networks in the 1960s and which is still very influential; and finally AdaTron learning, a refinement of the Perceptron algorithm with surprising dynamical features.

Universal costs of learning and a general time-energy-information trade-off

Our results for learning in neural networks immediately raise two questions. First, one may of course ask whether the total entropy production is a universal bound for learning for a reasonable class of learning problems and algorithms beyond the perceptron and supervised learning. Second, it is intriguing to wonder about the explicit role of time: although the inequalities derived in the previous chapters hold at all times, time does not explicitly enter the results. This raises the question whether there is a general trade-off between dissipation, information and learning duration.

In the final chapter of this thesis, we answer both these questions in the affirmative. We consider a very general inference problem, which encompasses the supervised learning problems of the previous sections, but also applies to unsupervised learning problems and architectures beyond the perceptron, such as multi-layer neural networks or Random Boltzmann machines, and allows for algorithms that employ feedback. Our approach is based on modelling the dynamics of all these systems using Bayesian networks. We prove an integral fluctuation theorem for a very broad class of learning problems and algorithms, including those with feedback. The second law that follows from this fluctuation theorem shows that the entropy production of the degrees of freedom of the learner is a universal bound for the information acquired during learning.

Second, we use the newly discovered thermodynamic uncertainty relation to derive a general relationship between energy, time and information that applies to learning. In this inequality, an undesirable quantity, the average entropy production of the whole system, bounds the product of two desirable quantities: the speed and the reliability of learning, which is the inverted variance of the total information acquired by the weights up to time t.

Concluding perspectives

The final analysis of the universal costs of learning completes the analysis of the stochastic thermodynamics of learning, which was the goal we set for ourselves at the outset.

Our work opens up numerous avenues for further research. Among them are the design of optimal algorithms from a thermodynamic perspective and the analysis of more complicated neural architectures. Looking towards quantum computation, it will be intriguing to study the trade-off between increased computational power and the inevitable dissipation increase that quantum coherence brings, at least in the linear response regime.

We have focused on the thermodynamic bounds on learning in this thesis, but dissipation is of course not the only thing that constraints the ability of neural networks and similar systems to compute. It has recently become clear that the behaviour of computational systems with respect to other constraints like the computation time and the amount of available data can be understood as phase transitions using statistical physics. It would be intriguing to see whether our thermodynamic bounds fit into this picture, and if so, where they can be found.

Zusammenfassung

Thermodynamik der Informationsvorbereitung

Informationsverarbeitung ist entscheidend für Lebewesen: Bakterien messen den externen Konzentrationsgradienten von Nahrungsmolekülen, um über ihre bevorzugte Schwimmrichtung zu entscheiden; Zellen geben präzise Kopien ihrer Erbinformation an ihre Nachfahren weiter, und neuronale Netzwerke sind in der Lage, komplexe Bewegungen zu planen und auszuführen. All diese Systeme sind bemerkenswert robust, obwohl sie starken thermischen Fluktuationen ausgesetzt sind, und sie können sie sehr präzise sein: *kinetic proofreading*, ein Mechanismus um genetische Information akkurat zu kopieren, erzielt eine Fehlerwahrscheinlichkeit von lediglich 3×10^{-8} pro Buchstaben des genetischen Codes.

Aus der Sicht der Thermodynamik sticht die Effizienz dieser Systeme hervor. *E.coli*, ein Bakterium, kann ein gegebenes Energie-Budget fast perfekt ausnutzen, um sein sensorisches Netzwerk an die Umgebung anzupassen, und das menschliche Gehirn benötigt lediglich 20 Watt. Diese Beobachtungen legen nahe, dass die Thermodynamik eine entscheidende Rolle bei der Evolution von informationsverarbeitenden Systemen in der Biologie gespielt hat.

Die statistische Physik fern vom Gleichgewicht hat in den letzten zwanzig Jahren große Fortschritte gemacht. Mit der stochastischen Thermodynamik ist dabei eine ausgereifte Rahmentheorie für das Studium von stark fluktuierenden Systemen fern vom Gleichgewicht entstanden, die auf der mathematischen Theorie der Markov-Prozesse basiert. Diese Theorie eignet sich insbesondere für die Analyse des Zusammenspiels zwischen Dissipation und Informationsverarbeitung. Im Bereich der Chemotaxis hat die stochastische Thermodynamik bereits wichtige Einblicke ermöglicht. Ermutigt von diesen Erfolgen wenden wir ihre Methoden in dieser Arbeit auf ein neues Problem an: das Lernen.

Informationsverarbeitung und Lernen in neuronalen Netzen

Das Ziel des Lernens ist es, aus bestehenden Datensätzen Modelle zu entwickeln und diese Modelle dann auf neue Fälle anzuwenden. Diese Prozesse werden in

Lebewesen von neuronalen Netzen geleistet, in denen eine große Anzahl von Nervenzellen, die sog. Neuronen, miteinander verbunden sind. Neuronen kommunizieren miteinander durch den Austausch von Aktionspotentialen. Dies sind kurze elektrische Signale, die über Synapsen von einem Neuron zum anderen übertragen werden. Ein einzelnes Neuron kann dabei als Schaltelement interpretiert werden: seine Eingabe ist die Aktivität aller Neuronen, mit denen es verbunden ist. Diese Neuronen können entweder ein Aktionspotential senden oder nicht. Die Ausgabe des Neurons ist wiederum binär: entweder versendet das Neuron ein weiteres Aktionspotential, oder es schweigt.

Aktionspotentiale werden über Synapsen zwischen den Neuronen übertragen. Die *Gewichtung* der Synapse, die Neuron A mit Neuron B verbindet, ist ein Maß dafür, wie sehr Neuron B von einem Aktionspotential beeinflusst wird, das von Neuron A gesendet wird. Je größer die Gewichtung, desto wahrscheinlicher ist es, dass B seinerseits mit einem Aktionspotential wreagiert. Die Gesamtheit der Gewichtungen in einem neuronalen Netzwerk regelt so, wie das Netzwerk Information verarbeitet. Die Änderung der Gewichtung wird allgemein als die neurophysiologische Grundlage von Lernen und Gedächtnis angesehen.

Unser Ansatz in dieser Arbeit ist das Studium von einfachen Modellen von Neuronen und insbesondere des bekannten *Perzeptron*-Modells, das von neurobiologischen Netzwerken inspiriert ist und eine große Rolle in der statistischen Physik und im maschinellen Lernen spielt. Das Perzeptron ist der Baustein, aus dem komplexere neuronale Netze gebaut werden können. Unser Fokus liegt auf der Dynamik der Gewichtungen des Perzeptrons. Die Änderung dieser Gewichtungen in endlicher Zeit in Gegenwart eines Wärmebads ist ein Nichtgleichgewichtsprozess, der thermodynamische Kosten in der Form von dissipierter freier Energie mit sich bringt. Die entscheidende Frage ist also die folgende: Wie viel freie Energie muss ein neuronales Netzwerk dissipieren, um zu lernen?

Modellfindung in neuronalen Netzen

Wir analysieren zunächst ein Perzeptron mit N Gewichtungen, welches ein Modell für einen Datensatz finden soll. Genauer gesagt werden dem Netzwerk eine Reihe von P Eingaben gegeben. Jede Eingabe hat ein wahres, binäres *Label*, das zufällig ausgewählt wird. Ein einzelnes Label ist dabei weder mit der dazugehörigen Eingabe noch mit den anderen Labels korreliert. Das Lernziel in diesem Problem ist, einen Satz von Gewichtungen zu finden, mit denen das Netzwerk das Label für so viele Eingaben wie möglich korrekt reproduzieren kann.

Wir zeigen in dieser Arbeit, dass die thermodynamischen Kosten bei der Modellfindung, also die totale Entropieproduktion der Gewichtungen, eine obere Schranke für die Information darstellen, die das Netzwerk aus dem Datensatz extrahieren kann. Wir messen diese Informationsmenge mit der gegenseitigen Information zwi-

schen den wahren Labeln der Eingaben und den Labeln, die das Netzwerk reproduziert. Unsere Ungleichung gilt für Netzwerke mit beliebig vielen Gewichtungen N und Datensätzen mit beliebig vielen Eingaben P.

Unsere Schranke ist scharf, da sie zum Beispiel saturiert wird von einem einfachen Netzwerk mit einer einzelnen Gewichtung, das das Label einer einzigen Eingabe lernt. Da die Zahl von Synapsen pro Neuron in realistischen Netzwerken durchaus in die Tausende gehen kann, studieren wir als ein weiteres Beispiel das Lernen im thermodynamischem Limes, wo wir die Zahl der Gewichtungen und die Zahl der zu lernenden Eingaben gegen unendlich gehen lassen, während das Verhältnis P/N von der Ordnung eins bleibt.

Verallgemeinerung zu unbekannten Eingaben

Das Perzeptron hat die bemerkenswerte Fähigkeit, von Beispielen lernen und das Gelernte auf neue Probleme übertragen zu können. Dies kann am besten für ein neuronales Netzwerk analysiert werden, das eine Regel lernt. Die Regeln, die wir lernen wollen, sind Boolesche Funktionen: sie weisen einer Eingabe ein binäres Label zu. Diese Regel wird von einem anderen neuronalen Netzwerk implementiert, dem Lehrer. Der Student muss die Form der Regel, also die Gewichtungen des Lehrers, aus einer Reihe von Beispielen für die Regel erschließen. Die Beispiele bestehen aus Eingabe-Label Paaren. Die Gewichtungen des Lehrers hingegen sind unzugänglich. Unser Interesse gilt der Fähigkeit des Studenten, nach einer Trainingsphase den Lehrer zu simulieren und das Label von Eingaben vorherzusagen, die das Netzwerk während der Trainingsphase nicht zu sehen bekam.

Dieses Problem ist insofern verschieden von der Modellfindung, die wir zunächst besprachen, als dass es in dem vorigen Problem keine sinnvolle Verallgemeinerung gibt: wenn alle Labels zufällig und unkorreliert zu den entsprechenden Eingaben oder den anderen Labels gezogen werden, dann sagt ein Datensatz mit Eingabe-Label Paaren nichts über das Label für eine weitere Eingabe aus. Bei einem Netzwerk, dass von einem Lehrer lernt, ist die Lage eine andere: hier sind die Labels alle vom Lehrer generiert und deswegen korreliert. Ein Datensatz mit einer gewissen Größe erlaubt so die Verallgemeinerung, d.h. die Vorhersage des Labels für Eingaben, die nicht Teil des Traningssatzes waren.

Wir messen die Fähigkeit des Studenten, die Regel von den Beispielen auf unbekannt Eingaben zu übertragen, mit der durchschnittlichen gegenseitigen Information zwischen dem wahren Label, bestimmt durch den Lehrer, und dem vorausgesagten Label des Studenten für eine beliebige Eingabe. Wir zeigen in dieser Arbeit, dass diese Information durch die durchschnittliche Entropieproduktion einer einzelnen Gewichtung im Netzwerk während des Lernens beschränkt ist.

Wir können diese Schranke mithilfe von Konzepten aus der stochastischen Thermodynamik für stationäre Zustände verschärfen. Dazu teilen wir die totale Entro-

pieproduktion in zwei Beiträge auf. Das Anlegen von äußeren Zwangsbedingungen, wie z.B. einer fluktuierenden Kraft auf die Gewichtungen, führt zur sogenannten *adiabatischen* Entropieproduktion. Das Treiben eines Systems von einem stationären in einen anderen stationären Zustand liefert einen einmaligen Beitrag zur totalen Entropieproduktion, die sog. *nicht-adiabatische* Entropieproduktion, deren Rate im stationären Zustand identisch null wird. Wir zeigen, dass die nicht-adiabatische Entropieproduktion ebenfalls eine obere Schranke für die gegenseitige Information zwischen dem wahren und dem vorhergesagten Label für eine beliebige Eingabe darstellt.

Unsere Ergebnisse gelten für eine große Klasse von Lernalgorithmen. Zur Illustration unserer Ergebnisse analysieren wir die Dynamik und die thermodynamische Effizienz dreier Algorithmen: Hebbsches Lernen, ein Algorithmus, der von neurobiologischen Studien in der Erinnerungsbildung inspiriert wurde; dem Perzeptron Algorithmus, dessen Entdeckung eine große Welle des Interesses an neuronalen Netzen in den 1960er Jahren auslöste und bis heute sehr einflussreich geblieben ist; und schließlich noch dem AdaTron Algorithmus, der eine Verfeinerung des Perzeptron Algorithmus darstellt und interessante dynamische Eigenschaften aufweist.

Die universelle Kosten des Lernens und die Beziehung zwischen Zeit, Energie und Information

Unsere bisherigen Ergebnisse werfen zwei Fragen auf. Zunächst gilt es, den ganzen Gültigkeitsbereich für thermodynamische Schranken an das Lernen zu bestimmen: können wir universelle Schranken für Lernprobleme und -algorithmen finden, die über das Perzeptron und das überwachte Lernen hinaus gehen? Weiterhin ist es interessant, über die Rolle der Zeit beim Lernen zu spekulieren. Die Ungleichungen, die wir bis jetzt besprachen, gelten zwar zu allen Zeiten, aber die Zeit taucht in den Formeln nicht explizit auf. Gibt es einen allgemeinen Zusammenhang zwischen Dissipation, Information und Lerndauer?

Im letzten Kapitel dieser Arbeit beantworten wir beide Fragen positiv. Wir analysieren das Lernproblem als allgemeines Inferenzproblem basierend auf kausalen Bayesianischen Netzwerken. Unser Modell beschreibt überwachtes und unüberwachtes Lernen, deckt viele verschiedenen Architekturen ab, darunter z.B. mehrlagige neuronale Netze oder Boltzmann Maschinen, und erlaubt die Verwendung von Lernalgorithmen, die auf Rückkopplung beruhen. Wir beweisen ein integrales Fluktuationstheorem für dieses Modell. Die daraus resultierende Form des zweiten Hauptsatzes zeigt, dass die Entropieproduktion in den Freiheitsgraden der Modelle eine universelle Schranke an die Menge an Information ist, die gelernt werden

kann.

Weiterhin nutzen wir die kürzlich entdeckte thermodynamische Unschärferelation um eine allgemeine Beziehung zwischen Dissipation, Zeit und Information für das Lernen herzuleiten. In der resultierenden Ungleichung beschränkt die Entropieproduktion des gesamten Systems, eine unerwünschte Größe, das Produkt aus zwei wünschenswerten Größen, nämlich dem Quadrat der Lerngeschwindigkeit sowie der Verlässlichkeit des Lernens, welches das Inverse von der Varianz der akquirierten Information ist.

Ausblick

Die abschließende Analyse der universellen thermodynamischen Kosten des Lernens vervollständigt die stochastische Thermodynamik des Lernens, deren Formulierung das Ziel dieser Arbeit war.

Unsere Arbeit wirft verschiedene Fragen zur weiteren Erforschung auf. Das Entwickeln von thermodynamisch optimalen Lernalgorithmen ist eine interessante, aber sicher auch sehr herausfordernde Aufgabe. Eine weiteres offenes Problem ist die Analyse der verschiedenen Kriterien, die komplexe, mehrlagige neuronale thermodynamisch effizient machen und welche dieser Kriterien in biologischen Netzwerken wiedergefunden werden.

Die Quantenphysik verspricht eine neue Generation von Computern, die unter Ausnutzung von Kohärenzen zwischen den Qubits gewaltige Zugewinne bei der Rechenleistung erreichen sollen. Kohärenzen zwischen Freiheitsgraden führen allerdings auch unausweichlich zu einer verminderten Effizienz von quantenmechanischen Wärmekraftmaschinen, zumindest im linearen Regime. Das Ausloten dieser gegensätzlichen Effekte – gesteigerte Rechenleistung bei gleichzeitig steigender Dissipation – stellt ein faszinierendes Forschungsprojekt dar.

Wir haben uns in dieser Arbeit auf die thermodynamischen Schranken des Lernens konzentriert, jedoch ist die Dissipation nicht die einzige Einschränkung für neuronale Netzwerke – man denke nur an die endliche Zeit, die für jede Berechnung zur Verfügung steht, oder die Größe der Datenmenge, die nötig ist, um erfolgreich Inferenzprobleme zu lösen. Das Verhalten von neuronalen Netzen hinsichtlich dieser Schranken kann mithilfe der statistischen Physik als Phasenübergang beschrieben werden. Es bleibt zu klären, ob unsere thermodynamischen Schranken ebenfalls in diesen Rahmen passen, und welche Rolle sie dort spielen.

Publications

Parts of this thesis have been or will be published in:

- S. Goldt and U. Seifert
  Universal costs of learning
  in preparation

- S. Goldt and U. Seifert
  Thermodynamic efficiency of learning a rule in neural networks
  New J. Phys. **19** 113001 (2017)

- S. Goldt and U. Seifert
  Stochastic Thermodynamics of Learning
  Phys. Rev. Lett. **11**, 11601 (2017)

- J. Fuchs, S. Goldt and U. Seifert
  Stochastic thermodynamics of resetting
  EPL **113** 60009 (2016)

1 Introduction

1.1 A new realm for thermodynamics

Information processing is ubiquitous in biological systems: Bacteria measure the gradients of the concentration of external nutrients to determine where to swim next [1]; cells pass faithful copies of their genetic information on to their progeny [2], and large neural networks are capable of performing complex motor control tasks [3]. These systems are all surprisingly robust, despite the fact that they are operating in noisy environments [1, 4], and they can be very precise: kinetic proofreading [5–7], a mechanism for high-fidelity copying of the genetic code, achieves an error probability of just 3×10^{-8} per letter [8].

From the perspective of thermodynamics, it is the efficiency of these processes which stands out: *E.coli*, a bacterium, is near-perfect in exploiting a given energy budget to adapt its sensory apparatus to changes in its environment [9], while the human brain consumes less than 20 watts [10]. These facts suggest thermodynamics played an important role in the evolutionary "design" of the chemical and neural networks that underlie biological computation. They motivate a detailed study of the thermodynamics of these systems, not least with an eye towards building nano-engines and improving the efficiency of current day super-computers, which operate in the megawatt range [11]. However, classical thermodynamics seems ill-suited for this endeavour on at least three levels: time, size, and driving.

Start with time. Classical thermodynamics is built around the notion of equilibrium states and reversible transitions between them, which take place infinitely slowly [12]. The brain, on the other hand, performs complex object recognition in about 150 ms [13].

Driving a system far from equilibrium is key to the success of biological information processing. The specificity of kinetic proofreading and other biochemical processes is better than the differences in equilibrium free energies would suggest [5]. This precision is possible because these systems operate far away from equilibrium, at the expense of continuously dissipating free energy. This regime is not covered by classical thermodynamics or non-equilibrium theories like linear response theory [14, 15].

Finally, classical thermodynamics applies to large systems. This is most clearly seen in statistical physics, which attempts to derive the thermodynamic laws for

large systems from the microscopic dynamics of its constituents, *e.g.* the single particles of gases [12, 16–18]. Key to this approach is the concept of a thermodynamic limit: as the system size increases (to infinity), a large assembly of microscopic particles can be characterised completely by a only few, macroscopic parameters like the temperature and pressure of a gas in a box. The evolution of these parameters is then studied in thermodynamics. Biological systems on the other hand are small: the reaction networks that underlie chemical sensing typically depend on the interplay of a number of molecules between 1 and 1000, to name but one example [19]. On this level, the dynamics are irreducibly stochastic and fluctuations, for example in the number of molecules, are important, but ignored by classical thermodynamics.

These observations make it clear that a lot of conceptual work was required to bring thermodynamics to the small, fluctuating systems far from equilibrium where biological computations take place. To this date, the most complete formulation of thermodynamics at this level is given by "stochastic thermodynamics" [20]. The remainder of this introduction will highlight some of its features with a focus on a number of model systems and experiments; a more technical discussion is given in Chapter 2.

1.2 Scaling down

Classical thermodynamics started in earnest with the quest for the ultimate limits of power generation by steam engines. This effort ultimately led to the formulation of equilibrium thermodynamics [23]. At the time, it was possible to ignore the role of thermal fluctuations because the size of the engines under consideration, such as the one shown in Fig. 1.1, rendered them negligible.

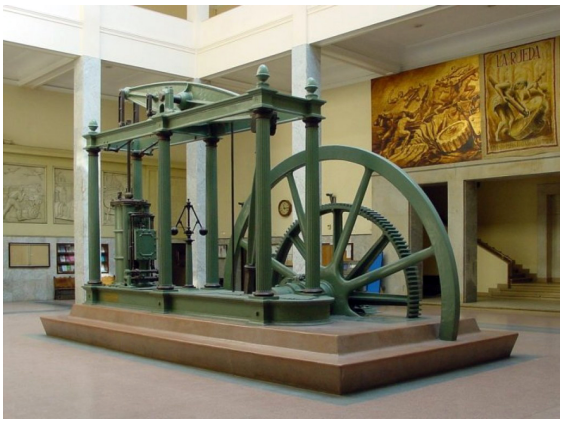
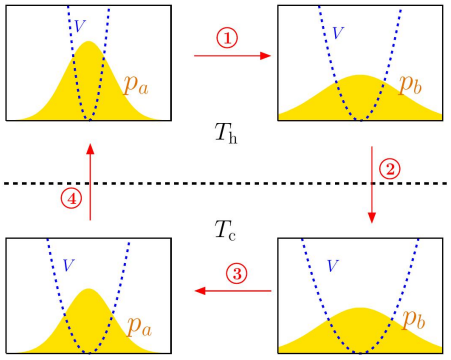
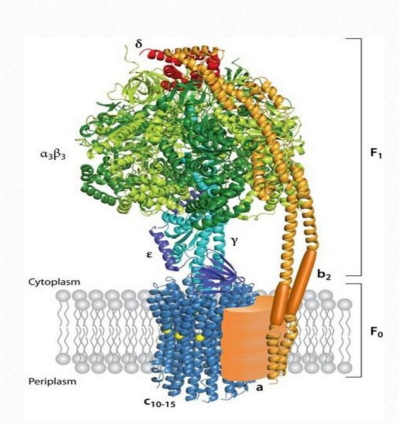


Figure 1.1 | A Watt steam engine built in 1859 in London. © 2004 by Nicolás Pérez. Reproduced with permission.

This changed with the rapid improvements in the manufacturing and controlling of mechanical devices down to the micro- or even nanometre length scale over the last three decades. They made it possible to design engines so small that the thermal fluctuations in the system are comparable to the energies of the system. The energy scale of these microscopic engines is typically on the order of few $k_{\mathrm{B}}T$, more than 20 orders of magnitude below the typical energy



a Schematic of a micrometre sized stochastic heat engine. Reprinted from [21].



b Cartoon of the enzyme ATP Synthase from X-ray crystallography. Reprinted from [2].

Figure 1.2 | Microscopic engines and the transformations they perform. **a** Recent advances in the manipulation of mechanical devices down to the nanometre length scale made it possible to build microscopic heat engines comprising just a single colloidal particle [22]. It performs a Stirling cycle to convert thermal energy into mechanical work. **b** The ATP synthase is a membrane-bound enzyme that exploits a chemical gradient to facilitate the synthesis of the energy storage molecule adenosine triphosphate (ATP).

scale of classical engines (here T is the temperature of the surrounding medium and k_B is Boltzmann's constant). This raises important conceptual questions for thermodynamics: can concepts like work and heat be meaningfully applied to small, fluctuating systems? Is there a consistent formulation of the second law of thermodynamics at this level? The short answer is yes: thermodynamics can be formulated consistently for small, fluctuating systems based on the mathematics of Markov processes. We discuss this in detail in Chapter 2.

Stochastic thermodynamics came full circle with the theoretical proposal [21] and experimental realisation [22] of a micro-metre sized heat engine. This engine consists of just a single colloidal particle, immersed in a fluid bath and controlled by an optical laser trap. It performs the equivalent of a classical Stirling cycle [12] on a length scale roughly nine orders of magnitude below the steam engines of Carnot's time, see Fig. 1.2a. Since then, the size of the smallest heat engines has shrunk even further, to machines using just a single atom [24, 25].

The appreciation of the molecular basis of life and the study of the molecular machinery that powers virtually all processes in cells opened up another avenue to use concepts of thermodynamics. A classic example is ATP synthase, a membrane-bound enzyme that exploits a chemical gradient to facilitate the syn-

thesis of adenosine triphosphate (ATP) [2]. This molecule has some particularly high-energy phosphate bonds which are used in a plethora of other chemical processes in the cell, making it the "energy currency" of molecular biology. ATP synthase is highly efficient, in spite of the large fluctuations in its surroundings [26–28]. Studying these molecules can give important insights for the design of artificial nano-machines [29].

But it is not just fluctuations and far from equilibrium dynamics which play an important role for the thermodynamics of small systems: so too does information, which we consider in the next section.

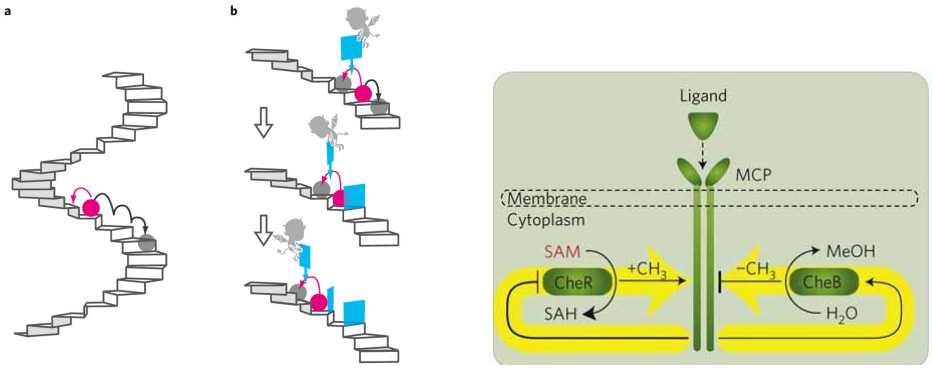
1.3 Information is physical

The concept of information was given precise mathematical meaning in Claude Shannon's remarkable 1948 paper [30], which created the field of information theory [31], see Sec. 2.3. Moreover, information is also physical [32]: it is stored and processed using physical hardware, so information processing has to adhere to the laws of physics. Thermodynamics in particular is closely intertwined with computation. This insight goes back to a time before computers, when J.C. Maxwell wondered whether the second law of thermodynamics would withstand the attack of an intelligent being, Maxwell's demon [17], who uses information to extract work from thermal fluctuations*.

A seminal observation was made by Landauer when he realised that the erasure of one bit of information leads to at least $k_{\mathrm{B}}T \ln 2$ of heat dissipation into the surrounding bath and hence has an associated thermodynamic cost [42]. This "Landauer principle" was recently experimentally verified [43, 44] using colloidal particles. These systems are examples of engines for which information is either a resource or the output. Figure 1.3a shows a concrete example of an engine that uses information as a resource and converts it into potential energy. This "demon" was realised in an ingenious experiment [41] involving a microscopic particle on a staircase-like potential with step heights on the order of $k_{\mathrm{B}}T$ (left of 1.3a). The key idea is to rectify the thermal fluctuations, which occasionally push the particle "up the staircase", by applying feedback control: as an upward jump is observed, a block is placed behind the particle to prevent downward jumps. Repeating this cycle allows the particle to climb up the stairs without energy injection, thereby converting the information gained by measurements into potential energy.

Information processing can easily be integrated into stochastic thermodynamics, yielding a framework to study the interplay of dissipation and information pro-

* This idea spawned a debate that has produced an enormous amount of literature and confusion and is still subject to active research and debate; for comprehensive reviews, see *e.g.* Refs.[33, 34], and for an analysis within stochastic thermodynamics Refs. [35–40].



a Schematic of information-to-work conversion using feedback. Reprinted from [41].

b The biochemical network underlying chemotaxis in *E.coli*. Reprinted from [9].

Figure 1.3 | Information can be a resource and an output of engines. **a** Information can be used as a resource and be transformed into potential energy, for example by using feedback to rectify the fluctuations in a colloidal particle to drive it up a spiral-staircase like potential. **b** Schematic of the chemical network that underlies chemotaxis in *E.coli*, a bacterium. In this network, chemical energy from ATP hydrolysis is used to monitor the concentration of nutrients in the surrounding medium, effectively converting potential into information.

cessing in small systems far from equilibrium. This integrated view paved the way for a whole new area of research: biological information processing.

1.4 Three steps of biological information processing

The sensory system of *E.coli* that we mentioned in the introduction measures the external concentration of nutrients to help the bacterium decide in which direction to swim next. The chemical network that underlies this sensing mechanism, schematically shown in Fig. 1.3b, can indeed be interpreted as an information engine. It is maintained in a non-equilibrium steady state by the continuous dissipation of free energy supplied by ATP molecules to precisely monitor the external concentration of nutrients. By converting potential energy into information, this engine performs the opposite operation of the demon in Fig. 1.3a.

Sensing has a history of interest from physicists going back at least to the seminal work of Berg and Purcell [45] searching for the fundamental limits on sensing imposed by physical laws. More recently, the application of stochastic thermodynamics has provided some intriguing results with regards to the physical limits on the speed, precision and thermodynamic cost of sensing [9, 46–58] and other biomolecular processes [59–63].

However, sensing is but the first of three steps of biological information processing. After acquiring sensory information about the external state of affairs, an organism has to build a model or a representation of the data is built to allow for efficient processing. Such a model is then the basis for the third and final step: processing previously unseen inputs by applying the model and making decisions based on the model's output, *i.e.* to "generalise" from past experience.

1.5 Aim and outline of this thesis

In this thesis, we study the last two steps of information processing, namely learning. In biological systems, learning and generalising is implemented in neural networks, where vast numbers of neurons communicate with each other via action potentials, the electric pulse used universally as the basic token of communication in neural systems [3]. We will interpret these neural networks as information processing engines and use (stochastic) thermodynamics to answer the two fundamental questions of thermodynamics: (1) what are the limits on neural networks imposed by thermodynamics and the Second Law in particular, and (2) how efficient are neural networks at learning, *i.e.* how much free energy do they have to dissipate in order to extract information from data or learn from examples?

This thesis is organised as follows. The next chapter gives a more technical introduction into stochastic thermodynamics, with a focus on continuous degrees of freedom and the thermodynamics of information processing. In chapter three, we introduce a key model for a neural network, the famous perceptron, and focus on the second step of information processing: building a model from data. In Chapter 4, we study the ability of the perceptron to learn a rule from examples provided by another neural network, called the teacher. We will derive thermodynamic bounds on the ability of such a network to learn from examples and to generalise its model to previously unseen problems. Finally, we derive an integral fluctuation theorem that imposes thermodynamic bounds on learning in neural network that applies to a wide range of learning problems and neural architectures in Chapter 5. We conclude this thesis by analysing the general trade-off between dissipation, speed and reliability of learning in neural networks. Some promising avenues for further research are outlined in Chapter 6.

2 Stochastic thermodynamics of information processing

Two new ideas reinvigorated non-equilibrium thermodynamics in 1997. First, K. Sekimoto realised that concepts from classical thermodynamics, in particular the first law, could be meaningfully applied to the stochastic trajectories of small, fluctuating systems [64, 65]. Separately, C. Jarzynski derived his famous non-equilibrium work equality [66, 67] which has the remarkable property of connecting the free energy differences between states, an equilibrium quantity, with a non-equilibrium quantity, namely the work spent to drive the system from one state to the other in *finite* time*. These discoveries triggered an ambitious research programme that has given rise to a framework called stochastic thermodynamics [72].

In this chapter, we will review some key concepts from stochastic thermodynamics. Given the enormous amount of activity over the last 20 years, our selection of topics here will necessarily be limited and guided by the requirements of the following chapters. For example, we will focus on continuous degrees of freedom with overdamped dynamics, although we stress that the concepts introduced in this chapter apply to any form of Markovian dynamics, with discrete systems governed by master equations being another important class. An introduction to the theory of (Markovian) stochastic processes from a physical point of view can be found in the books by van Kampen [73], Gardiner [74] and Risken [75], while a comprehensive review of stochastic thermodynamics was given by Seifert [20]; see also Refs. [40, 76–79] and for an overview over most recent developments, see Ref. [80].

In the following, we will start with the dynamics and thermodynamics of a colloidal particle on the level of single trajectories and on the ensemble level and discuss their relation. We then analyse the interplay of dissipation and information processing in N-particle systems and comment on the special features of non-equilibrium steady states.

* The Jarzynski relation is indeed an example of what is now called a fluctuation theorem. The first example of such a relation, which would now be called the steady-state fluctuation theorem for entropy production, was discovered in numerical simulations of shear-driven flow and justified heuristically by Evans, Cohen and Morriss [68]. It was later proven for different dynamics and system classes [69–71].

2.1 Thermodynamics on single trajectories

We consider the paradigmatic case of a colloidal particle immersed in a heat bath, which is in equilibrium at temperature T. We will assume that the dynamics of the particle are overdamped. It is then fully described by its position $x(t)$, which we interpret as the state of the random variable $X(t)$. On the level of single trajectories, the particle's position obeys the Itô stochastic differential equation [74] (SDE) for an infinitesimal time step $\mathrm{d}t$,

$$\mathrm{d}x(t) \equiv x(t+\mathrm{d}t) - x(t) = \mu F(x(t),\lambda)\,\mathrm{d}t + \mathrm{d}W(t) \tag{2.1}$$

where μ is the mobility of the particle and the Wiener process $\mathrm{d}W(t)$ is a random process whose increments are normally distributed with mean 0 and variance $2D\,\mathrm{d}t$. The diffusion constant D has to fulfil the fluctuation-dissipation relation

$$D = T\mu \tag{2.2}$$

to ensure thermodynamic consistency of the Markovian dynamics [20]. In other words, we assume that the degrees of freedom that make up the thermal bath are always in equilibrium. Here and throughout this thesis, we set Boltzmann's constant $k_\mathrm{B} = 1$ to make entropy dimensionless without loss of generality*.

Augmented with an initial condition $x(0) = x_0$, this SDE generates individual trajectories, depending on the realisation of the Wiener process, which models the thermal fluctuations of the bath. The force acting on the particle can arise due to a potential $V(x,\lambda)$ or a non-conservative force† $f(x,\lambda)$, so

$$F(x,\lambda) = -\partial_x V(x,\lambda) + f(x,\lambda) \tag{2.3}$$

where $\partial_x \equiv \partial/\partial_x$. Both contributions may vary in time via the control parameter $\lambda(t)$. The Langevin equation for the position of the particle $x(t)$, which is more prevalent in the physical literature, reads

$$\dot{x}(t) = \mu F(x,\lambda) + \zeta(t) \tag{2.4}$$

where the noise term $\zeta(t)$ is related to the Wiener process by $\mathrm{d}W(t) = \int_t^{t+\mathrm{d}t} \mathrm{d}t'\zeta(t')$ and obeys

$$\langle \zeta(t) \rangle = 0, \qquad \langle \zeta(t)\zeta(t') \rangle = 2D\delta(t-t') \tag{2.5}$$

* For a system with discrete states and master equation dynamics, this is equivalent to imposing a local-detailed balance condition on the ratio of transition rates between any two states.

† Although any force can always be written in terms of a potential in one dimension, we maintain this separation because already in two dimensions, there are genuinely non-conservative forces and the distinction will be useful from a conceptual point of view in the following.

We will use angled brackets $\langle\cdot\rangle$ to indicate averages over thermal noise throughout this thesis, unless indicated otherwise. However, we note that only the SDE is mathematically sound [73, 74] due to the irregularity of the noise term $\zeta(t)$ in the Langevin equation. This will be important when we try to identify thermodynamic quantities on the level of single trajectories.

2.1.1 Energetics and the first law

Sekimoto realised that the equation of motion (2.1) can be interpreted as a formulation of the first law [64, 65, 78]. The most transparent way to see this is to first consider the change in potential energy of a particle that is only subject to a conservative force, *i.e.* $f = 0$. We then have

$$\mathrm{d}V \equiv \mathrm{d}\lambda \left.\frac{\partial V}{\partial \lambda}\right|_{x(t),\lambda(t)} + \mathrm{d}x(t) \left.\frac{\partial V}{\partial x}\right|_{x(t),\lambda(t)} + \mathrm{d}x^2(t)\frac{1}{2}\left.\frac{\partial^2 V}{\partial x^2}\right|_{x(t),\lambda(t)} \tag{2.6}$$

where we have to make sure that we expand the potential (or any function of $x(t)$) to second order in $x(t)$ to obtain all the changes which are first order in time due to the presence of the Wiener process. We can identify the first term in Eq. (2.6) as the work done on the particle, which we will take to be positive,

$$đw \equiv \mathrm{d}\lambda \left.\frac{\partial V}{\partial \lambda}\right|_{x(t),\lambda(t)} . \tag{2.7}$$

We will use small letters to denote quantities pertaining to individual trajectories throughout. The heat dissipated into the environment due to friction [64, 81] is the negative of the second term,

$$đq \equiv -\mathrm{d}x(t) \left.\frac{\partial V}{\partial x}\right|_{x(t),\lambda(t)} - \mathrm{d}x^2(t)\frac{1}{2}\left.\frac{\partial^2 V}{\partial x^2}\right|_{x(t),\lambda(t)} \tag{2.8}$$

$$= -\mathrm{d}x(t)\frac{1}{2}\left[\left.\frac{\partial V}{\partial x}\right|_{x(t),\lambda(t)} + \left.\frac{\partial^2 V}{\partial x^2}\right|_{x(t+\mathrm{d}t),\lambda(t)}\right] \tag{2.9}$$

$$\equiv -\mathrm{d}x(t) \circ \left.\frac{\partial V}{\partial x}\right|_{x(t),\lambda(t)} = \mathrm{d}x(t) \circ F(x(t),\lambda(t)) \tag{2.10}$$

where we have defined the Stratonovich product $\circ$ which corresponds to the "midpoint" rule for evaluating stochastic integrals and for which the ordinary rules of calculus apply [74]. We note that the identification of the heat dissipated into the reservoir as the Stratonovich product of total force on the particle and incremental particle displacement is also valid with $f \neq 0$ and for underdamped dynamics.

2.1.2 Entropy and the second law

Entropy can also be identified on the level of single trajectory. This quantity has two contributions. First, the heat dissipated into the medium that we just identified is associated with an increase of medium entropy

$$\Delta s^{\mathrm{m}}[x(t)] \equiv q[x(t)]/T \tag{2.11}$$

where $q[x(t)]$ is the total heat dissipated along the trajectory $x(t)$. The second contribution is identified as the stochastic entropy [82] or system entropy,

$$s(t) \equiv -\ln p[x(t), t] \tag{2.12}$$

which is evaluated by first solving the ensemble dynamics for $p(x,t)$ and then evaluating the distribution for the given time at the trajectory-dependent position.

The distribution $p(x,t)$ for particle positions generated by the Langevin dynamics (2.1) obeys the corresponding Fokker-Planck equation [75] (FPE)

$$\partial_t p(x,t) = -\partial_x j(x,t) \tag{2.13}$$

$$\equiv -\partial_x \left[\mu F(x, \lambda(t)) p(x,t) - D\partial_x p(x,t)\right] \tag{2.14}$$

where we have defined the probability current $j(x,t)$. The FPE can be directly derived from the SDE (2.1) using Itô's rule [74]. It requires an initial condition $p(x,0)$ and a specification of the boundary conditions; we will usually assume natural boundary conditions, where $p(x,t) \to 0$ as $x \to \pm\infty$ and likewise for $j(x,t)$. The stochastic entropy of the particle at any time hence depends on the ensemble from which the trajectory was taken.

The total entropy production of the particle along a trajectory thus reads

$$\Delta s^{\mathrm{tot}} \equiv \Delta s^{\mathrm{m}}[x(t)] + \Delta s = \Delta s^{\mathrm{m}}[x(t)] - \ln p[x(t)] + \ln p[x(0)] \tag{2.15}$$

and obeys an integral fluctuation theorem for arbitrary time- dependent driving $\lambda(t)$, arbitrary length t of the process and arbitrary initial condition $p(x,0)$ [82]

$$\left\langle e^{-\Delta s^{\mathrm{tot}}} \right\rangle = 1 \tag{2.16}$$

where the average is taken with respect to $p(x,t)$. Applying Jensen's inequality [31] to (2.16) and introducing capital letters to denote ensemble quantities, we find that

$$\Delta S^{\mathrm{tot}} \equiv \left\langle \Delta s^{\mathrm{tot}} \right\rangle \geq 0 \tag{2.17}$$

which is reminiscent of another well-known inequality in thermodynamics. However, this result should not be considered a "proof" of the second law, just as much

as negative values of Δs^{tot} are no "violation" of the second law: both these results derive from the fact that we introduced a fundamental irreversibility from the beginning by our choice of dynamics, and classical thermodynamics is of course silent about issues which lie outside of its range of validity – *e.g.* small, strongly fluctuating systems. Instead, the integral fluctuation theorem is a refinement of the second law to arbitrary moments of the total entropy production.

2.2 Ensemble thermodynamics

The average $\langle \Delta s^{\mathrm{tot}} \rangle$ that entered our statement of the second law in Eq. (2.17) is an example of an ensemble average over trajectories. They should coincide with the quantities that can be independently derived on the ensemble level from a Fokker-Planck or master equation. We will indicate ensemble averages over noise ensembles using angled brackets, $\langle \cdot \rangle$, unless indicated otherwise, and denote all quantities on the ensemble level with capital letters, *e.g.* $\Delta S^{\mathrm{tot}} = \langle \Delta s^{\mathrm{tot}} \rangle$.

The average energy of the overdamped system is given by

$$E(t) = \int \mathrm{d}\boldsymbol{x}\; p(\boldsymbol{x}, t) V(\boldsymbol{x}, t) \tag{2.18}$$

The average of the stochastic entropy (2.12),

$$S(X) = -\int \mathrm{d}x\; p(x) \ln p(x) \tag{2.19}$$

is readily identified as the Shannon entropy [30] of the system, which is a measure of the average uncertainty an observer has about the current state of x. We will come back to the subject of information theory in Section 2.3.

Using the Fokker-Planck equation (2.13) and integrating by parts, we find

$$\begin{aligned} \partial_t S(X) &= \int \mathrm{d}x\; \partial_t p(x,t) \ln p(x,t) \\ &= \int \mathrm{d}x\; j(x,t) \frac{\partial_x p(x,t)}{p(x,t)} \\ &= \int \mathrm{d}x\; j(x,t) \left(\frac{j(x,t)}{D p(x,t)} - F(x,\lambda)/T \right) \end{aligned} \tag{2.20}$$

where we have used the definition of the probability current, Eq. (2.14) for the last equality. We can identify the average rate of entropy production in the medium,

$$\dot{S}^{\mathrm{m}} = \int \mathrm{d}x\; j(x,t) F(x,t)/T \tag{2.21}$$

and the average rate of total entropy production

$$\dot{S}^{\mathrm{tot}} = \partial_t S + \dot{S}^{\mathrm{m}} = \int \mathrm{d}x \, \frac{j(x,t)^2}{Dp(x,t)} \geq 0 \tag{2.22}$$

which is evidently positive, with equality in equilibrium only. Here and throughout this thesis, we will use the overdot notation to distinguish rates from the time derivatives of state function, such as the Shannon entropy $S(X)$.

2.2.1 Consistency of trajectory and ensemble quantities

The consistency of the framework that we have introduced in the last two sections requires that the quantities we introduced on the level of single trajectories are equal to the quantities that were derived for Fokker-Planck dynamics on the ensemble level. The most straightforward way to show this is by averaging the trajectory-dependent quantities from Section 2.1. While the averages of quantities that involve only the position x are easily done using the distribution $p(x,t)$, for example

$$\langle s(t) \rangle = \langle -\ln p(x(t),t) \rangle = -\int \mathrm{d}x \; p(x,t) \ln p(x,t) = S(X,t) \tag{2.23}$$

as expected. Averages over functions of derivatives like $\dot{x}$ are more delicate, however, due to the intricacies of stochastic calculus that we alluded to before. However, a short calculation [20, 82] shows that for any function $f(x)$

$$\langle f(x)\dot{x} \rangle = \left\langle f(x) \frac{j(x,t)}{p(x,t)} \right\rangle = \int \mathrm{d}x f(x) j(x,t). \tag{2.24}$$

2.3 A quick primer on information theory

We have briefly mentioned that the average of the stochastic entropy is given by the Shannon entropy of the system, which is the quantity at the heart of information theory. Let us briefly expand on this point and introduce a number of key concepts from information theory which will prove indispensable for our analysis of information processing in neural networks and beyond. We will be brief here and only state a number of definitions and give some intuition on the most important quantities. For modern expositions of the subject, consult the encyclopedic book by Cover and Thomas [31] or the original exposition of MacKay [8].

Information theory was started in a very influential paper by C. Shannon in 1948 [30]. His aim was to quantify the information content $s(x)$ of a signal $x \in \mathcal{X}$, which he defined as

$$s(x) \equiv -\ln p(x). \tag{2.25}$$

It is measured in bits and intuitively assigns the highest information content to the event with the lowest probability. For a continuous random variable $x \in \mathcal{X}$ with support $\mathcal{X}$, the average of the surprise yields the Shannon entropy as seen before, which we repeat here for completeness,*

$$S(X) \equiv \langle s(x) \rangle = -\int_{\mathcal{X}} \mathrm{d}x \; p(x) \ln p(x) \geq 0 \tag{2.26}$$

where the integration runs over the support of X and will from now on be understood implicitly†. The Shannon entropy is hence the average uncertainty one has about the value or a random variable X. Another interpretation that is particularly suitable for discrete random variables and computing entropies using the logarithm with base 2 is that the Shannon entropy gives the average number of yes-or-no questions about a system that need to be answered in order to correctly identify x.

For two random variables X and Y with joint distribution, the definition (2.26) also applies to their joint distribution: $S(X,Y) \equiv -\langle \ln p(x,y) \rangle$, where the average is taken over the joint distribution $p(x,y)$. The conditional Shannon entropy, which measures the average uncertainty an observer has about the value of X *given* knowledge about the state of Y, is then defined as

$$S(X|Y) \equiv -\int \mathrm{d}x \, \mathrm{d}y \; p(x,y) \ln p(x|y) \leq S(X) \tag{2.27}$$

where the conditional probability $p(x|y) \equiv p(x,y)/p(y)$. A natural measure for the correlations between X and Y is then the amount by which knowledge about Y reduces the uncertainty about X compared to the *a priori* uncertainty about X, $S(X)$. We hence define the mutual information as

$$I(X:Y) \equiv S(X) - S(X|Y) \tag{2.28}$$

$$= \int \mathrm{d}x \, \mathrm{d}y \; p(x,y) \ln \frac{p(x,y)}{p(x)p(y)} \geq 0 \tag{2.29}$$

where the last inequality follows from the log-sum inequality. From the last form, we immediately see that the mutual information is symmetric in its arguments $I(X:Y) = I(Y:X)$. We note that X and Y can be any number of variables.

* Legend has it that it was the formal equivalence of this expression to the Gibbs entropy that led von Neumann to suggest the name "entropy" to Shannon as he was looking for a good name for this quantity; that way, von Neumann quipped, Shannon would always have the upper hand in any argument – because no one knows what entropy actually is [83].

† We note that for continuous random variables, the integral (2.26) might not exist. The Shannon entropy of a continuous variable may also become negative; in any case, the mutual information (2.28) is well defined.

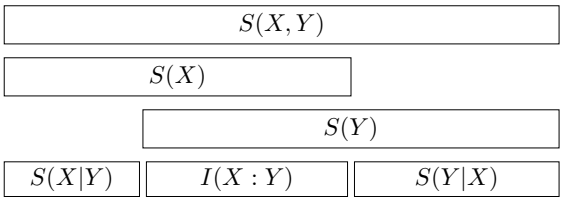


Figure 2.1 | Relationship between Shannon entropies and mutual information. The definitions of all quantities are given in the text.

The conditional mutual information of X and Y given another random variable Z is defined as

$$I(X:Y\,|\,Z) \equiv S(X|Z) - S(X|Y,Z). \tag{2.30}$$

We summarise the relationship between joint entropy, conditional entropy and mutual information in Fig. 2.1 with a word of caution: this graphical approach cannot be applied to Venn-like diagrams, which are sometimes used; see for example Ref. [8] for a discussion.

Finally, we introduce the Kullback-Leibler divergence between to distributions with the same support $p(x)$ and $q(x)$:

$$D\left[p(x) \parallel q(x)\right] \equiv \int \mathrm{d}x\; p(x) \ln \frac{p(x)}{q(x)} \geq 0. \tag{2.31}$$

The Kullback-Leibler divergence is also sometimes referred to as a distance, but we emphasise that is not symmetric in its arguments, $D\left[p \parallel q\right] \neq D\left[q \parallel p\right]$, and that it does not, in general, obey the triangle inequality. The mutual information is an example of a Kullback-Leibler divergence: $I(X:Y) = D\left[p(x,y) \parallel p(x)p(y)\right]$.

2.4 Multipartite dynamics and information processing

Let us now extend this discussion to a system with N interacting degrees of freedom, $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ that are controlled by a number of possibly time-dependent control parameters $\boldsymbol{\lambda}(t)$. The system has states $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)$. Every degree of freedom or subsystem X_n is connected individually to a distinct heat bath at constant temperature T_n. This allows us to uniquely identify the heat flows in and out of each system and makes the dynamics of the system multipartite [38, 39, 84], meaning that the fluctuations in any subsystem are uncorrelated from the fluctuations in all the others. For concreteness, we shall consider a number of colloidal particles, such that $\boldsymbol{X} \in \mathbb{R}^N$ and the dynamics are overdamped,

but the ideas apply to any form of Markovian dynamics [73, 74]. Each particle then obeys a Langevin equation

$$\dot{x}_n(t) = \mu_n F_n(\boldsymbol{x}(t), \boldsymbol{\lambda}(t)) + \zeta_n(t). \tag{2.32}$$

As before, the force $F_n(\cdot)$ on a particle can arise due to a potential $V(\boldsymbol{x}, \boldsymbol{\lambda})$ and/or a non-conservative external force $f_n(\boldsymbol{x}, \boldsymbol{\lambda})$. For thermodynamic consistency, the Gaussian noise $\zeta_n(t)$ has correlations which obey the fluctuation-dissipation theorem and are uncorrelated across subsystems: $\langle \zeta_n(t)\zeta_m(t')\rangle = 2D_n\delta_{nm}\delta(t-t')$ with $D_n \equiv \mu_n T_n$ [20].

On the ensemble level, the system is fully described by its time-dependent probability distribution $p(\boldsymbol{x}, t)$, which obeys a multi-dimensional Fokker-Planck equation [74]. Since we are considering time-dependent dynamics, we note that all the quantities that we introduce in the remainder of this chapter will be time-dependent, even if we sometimes omit the time parameter t for clarity of notation. Thanks to the multipartite assumption, the total probability current $j(\boldsymbol{x}; t)$ splits into a distinct contribution due to each subsystem. We can thus write

$$\partial_t p(\boldsymbol{x}, t) = -\sum_{n=1}^{N} \partial_n j_n(\boldsymbol{x}, t) \tag{2.33}$$

with $\partial_n \equiv \partial/\partial x_n$ and $j_n(\boldsymbol{x}, t)$ is the probability current due to the fluctuations of the nth subsystem,

$$j_n(\boldsymbol{x}, t) = \mu_n F_n(\boldsymbol{x}, \boldsymbol{\lambda}) p(\boldsymbol{x}, t) - \mu_n T_n \partial_n p(\boldsymbol{x}, t) \tag{2.34}$$

We can write down the second law for the entire system

$$\dot{S}^{\mathrm{tot}} = \partial_t S(\boldsymbol{X}) + \dot{S}^{\mathrm{m}} \tag{2.35}$$

where $S(\boldsymbol{X})$ is now the Shannon entropy of the full distribution $p(\boldsymbol{x})$. Throughout this section, we will use the overdot notation to distinguish rates from the change of state functions, such as $S(\boldsymbol{X})$. The total entropy production can be split into a separate contribution due to the fluctuations of each subsystem X_n due to the multipartite property of the dynamics (2.33):

$$\partial_t S(\boldsymbol{X}) = \sum_n \dot{S}_n(\boldsymbol{X}) = -\sum_n \int \mathrm{d}\boldsymbol{x}\; j_n(\boldsymbol{x}, t) \partial_n \ln p(\boldsymbol{x}, t). \tag{2.36}$$

The rate of increase of medium entropy $\dot{S}^{\mathrm{m}}$ due to the heat dissipated into each environment can also be split into a contribution due to each subsystem,

$$\dot{S}^{\mathrm{m}} = \sum_n \frac{\dot{Q}_n}{T_n} = \sum_n \int \mathrm{d}\boldsymbol{x}\; j_n(\boldsymbol{x}, t) F_n(\boldsymbol{x}, \boldsymbol{\lambda}(t))/T_n. \tag{2.37}$$

Following the lines of Eq. (2.20), we can write the total rate of entropy production compactly as

$$\dot{S}^{\mathrm{tot}} = \sum_n \dot{S}_n^{\mathrm{tot}} = \sum_n \left[\dot{S}_n(\boldsymbol{X}) + \dot{S}_n^{\mathrm{m}}\right] = \sum_n \int \mathrm{d}\boldsymbol{x}\, \frac{j_n(\boldsymbol{x},t)^2}{D_n p(\boldsymbol{x})} \tag{2.38}$$

from which it follows immediately that the second law also holds on the level of each subsystem individually:

$$\dot{S}_n^{\mathrm{tot}} = \dot{S}_n(\boldsymbol{X}) + \dot{S}_n^{\mathrm{m}} = \int \mathrm{d}\boldsymbol{x}\, \frac{j_n(\boldsymbol{x},t)^2}{D_n p(\boldsymbol{x})} \geq 0. \tag{2.39}$$

2.4.1 Disentangling the flow of information

We can gain insight into the information processing of this system by substituting $p(\boldsymbol{x},t) = p(x_n,t)p(\overline{\boldsymbol{x}},t|x_n)$, with $\overline{\boldsymbol{x}}_n \equiv (x_1,\ldots,x_{n-1},x_{n+1},\ldots,x_N)$, in the expression for $\dot{S}_n$, Eq. (2.36). The expression thus separates it into two parts: first, the change of Shannon entropy of the marginalised distribution $p(x_n)$,

$$\dot{S}_n(X_n) = -\int \mathrm{d}\boldsymbol{x}\; j_n(\boldsymbol{x},t)\partial_n \ln p(x_n,t) = \partial_t S(X_n), \tag{2.40}$$

where the last equality follows from the fact that an entropy change of the marginalised distribution $p(x_n)$ can only come from the dynamics of x_n. The second part reads

$$l_n(x_n;\overline{\boldsymbol{x}}) = \int \mathrm{d}\boldsymbol{x}\; j_n(\boldsymbol{x},t)\partial_n \ln p(\overline{\boldsymbol{x}}_n|x_n) \tag{2.41}$$

and is called the learning rate [38] or information flow [39, 85]. This thermodynamic learning rate* is a thermodynamically consistent measure of how much the dynamics of x_n change the mutual information $I(X_n : \overline{\boldsymbol{X}}_n)$. The second law (2.39) for the nth subsystem hence becomes

$$\dot{S}_n^{\mathrm{tot}} = \partial_t S(X_n) + \dot{Q}_n/T_n - l_n(X_n;\overline{\boldsymbol{X}}) \geq 0 \tag{2.42}$$

This form of the second law hints at the way that information can be a resource for information-driven engines, *cf.* Sec. 1.3. The apparent entropy production of the nth particle, which is the entropy production we would assign to it if it were alone, is

$$\sigma_n = \partial_t S(X_n) + \dot{Q}_n/T_n, \tag{2.43}$$

* We emphasise that this learning rate l_n arises from thermodynamic considerations and has nothing to do with the learning rate ν that goes into the definition of learning algorithms, for example in Eq. (3.12) or Eq. (4.8). To avoid confusion, we will refer to l_n as the thermodynamic learning rate whenever it isn't clear from context which quantity we refer to.

and we would require it to be positive by the second law. Instead, the presence of the learning rate (2.41) allows that the entropy of the nth system and its environment *decreases*, $\sigma_n < 0$, by using the learning rate as a resource that maintains the positivity of the total entropy production rate $\dot{S}_n^{\text{tot}}$.

We can further refine the second law (2.42) by exploiting the causal structure of the dynamics, as was recently suggested by Horowitz [84]. The subsystem X_n interacts directly only with its "neighbours", *i.e.* those degrees of freedom $\boldsymbol{X}_{\mathcal{C}_n} \subseteq \overline{\boldsymbol{X}}_n$ that enter the force acting on it, *i.e.* $F_n = F_n\left(x_n, \boldsymbol{x}_{\mathcal{C}_n}, \boldsymbol{\lambda}(t)\right)$. Keeping this in mind, we use the chain rule for mutual information [31] to write

$$I(X_n : \overline{\boldsymbol{X}}_n) = I(X_n : \boldsymbol{X}_{\mathcal{C}_n}) + I(X_n : \overline{\boldsymbol{X}}_{\mathcal{C}_n} \,|\, \boldsymbol{X}_{\mathcal{C}_n}), \tag{2.44}$$

with $\boldsymbol{X}_{\mathcal{C}_n} \cup \overline{\boldsymbol{X}}_{\mathcal{C}_n} = \overline{\boldsymbol{X}}_n$ and the conditional mutual information

$$\begin{aligned} I(X_n : \overline{\boldsymbol{X}}_{\mathcal{C}_n} \,|\, \boldsymbol{X}_{\mathcal{C}_n}) =& S(X_n|\boldsymbol{X}_{\mathcal{C}_n}) - S(X_n|\overline{\boldsymbol{X}}_{\mathcal{C}_n}, \boldsymbol{X}_{\mathcal{C}_n}) && (2.45)\\ =& \int \mathrm{d}\boldsymbol{x}\; p(\boldsymbol{x}, t) \ln \left(\frac{p(\boldsymbol{x}, t) p(\boldsymbol{x}_{\mathcal{C}_n})}{p(x_n, \boldsymbol{x}_{\mathcal{C}_n}) p(\overline{\boldsymbol{x}}_{\mathcal{C}_n}, \boldsymbol{x}_{\mathcal{C}_n})} \right). && (2.46) \end{aligned}$$

Accordingly, we split the thermodynamic learning rate (2.41) into a thermodynamic learning rate of x_n with the degrees of freedom that it directly interacts with, *i.e.* the $\boldsymbol{X}_{\mathcal{C}_n}$,

$$l_n(X_n; \boldsymbol{X}_{\mathcal{C}_n}) = \int \mathrm{d}\boldsymbol{x}\; j_n(\boldsymbol{x}, t) \partial_n \ln p(\boldsymbol{x}_{\mathcal{C}_n}|x_n), \tag{2.47}$$

and a thermodynamic learning rate with the other subsystems given its neighbours,

$$l_n(x_n; \overline{\boldsymbol{X}}_{\mathcal{C}_n}|\boldsymbol{X}_{\mathcal{C}_n}) = \int \mathrm{d}\boldsymbol{x}\; j_n(\boldsymbol{x}, t) \partial_n \ln\left(\cdot\right). \tag{2.48}$$

By simple integration, one can derive the following second-law like inequality,

$$\partial_t S(x_n) + \dot{Q}_n - l_n(x_n; \boldsymbol{x}_{\mathcal{C}_n}) \geq 0, \tag{2.49}$$

which now includes the refined thermodynamic learning rate (2.47). We finally note for completeness that [84]

$$l_n(x_n; \overline{\boldsymbol{X}}_{\mathcal{C}_n}|\boldsymbol{X}_{\mathcal{C}_n}) \leq 0. \tag{2.50}$$

We will make extensive use of this framework to analyse the information processing in the neural networks that we study.

2.5 Steady state thermodynamics

Finally, let us consider a particular class of states for which there exist particularly strong results. A steady state is any state where for fixed values of the control parameters $\boldsymbol{\lambda}$, $\partial_t p^{\mathrm{s}}(\boldsymbol{x}) = 0$. Our definition hence includes both genuine equilibrium states, where all currents vanish, $j_n^{\mathrm{eq}}(\boldsymbol{x}) = 0$, and non-equilibrium steady states (NESS), where at least some of the currents are non-zero.

Of course, the second law still applies to steady states, but is not sharp anymore: $\Delta S(x) \sim 1$ while steady-states have a non-vanishing rate of entropy production and hence $\Delta Q \sim t$. We will be interested not just in the cost of maintaining the steady state for a given set of control parameters, but also in the thermodynamic costs of driving the system from one steady state to another (one of which may be equilibrium).

Steady-state thermodynamics [86–88] provides a framework to disentangle the housekeeping heat, dissipated to maintain a given steady state, and the excess heat, which arises from transitions between steady states. It builds on stochastic thermodynamics and it has been experimentally verified [89]. In this section, we briefly illustrate some of its key concepts for a particle on a ring and we discuss the multipartite case.

2.5.1 A particle on a ring

We will introduce the different entropy production rates of steady-state thermodynamics by considering the paradigmatic example of a single colloidal particle with position x on a ring which is dragged by a constant force λ.

For every value of λ, there is a well-defined, unique [73] steady-state distribution $p^{\mathrm{s}}(x, \lambda)$ with probability current $j^{\mathrm{s}}(x, \lambda) \neq 0$ and a particular rate of thermodynamic entropy production $\dot{S}^{\mathrm{m}} \geq 0$. Let us drop the arguments from the currents and distribution for the remainder of this section and rewrite the total entropy production of the system using the shorthands

$$\dot{S}^{\mathrm{tot}} = \int \mathrm{d}x \, \frac{j^2}{Dp} = \int \mathrm{d}x \, \frac{p}{D} \left(\frac{j}{p} - \frac{j^{\mathrm{s}}}{p^{\mathrm{s}}} + \frac{j^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \tag{2.51}$$

where j^{s} is understood to be the probability current if the system were to relax with λ fixed at $\lambda(t)$. From this expression, we can identify two contributions to the total entropy production [90], namely the non-adiabatic entropy production

$$\dot{S}^{\mathrm{na}} \equiv \int \mathrm{d}x \, \frac{p}{D} \left(\frac{j}{p} - \frac{j^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \geq 0 \tag{2.52}$$

and the adiabatic entropy production

$$\dot{S}^{\mathrm{a}} \equiv \int \mathrm{d}x \, \frac{p}{D} \left(\frac{j^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \geq 0 \tag{2.53}$$

which are both evidently positive. Once we have reached the steady state, $\dot{S}^{\mathrm{a}} = \dot{S}^{\mathrm{tot}}$ and $\dot{S}^{\mathrm{na}} = 0$. The cross-term from the binomial in Eq. (2.51) cancels, which is easily seen by partial integration and using that $\partial_x j^{\mathrm{s}} = 0$, hence we have

$$\dot{S}^{\mathrm{tot}} = \dot{S}^{\mathrm{a}} + \dot{S}^{\mathrm{na}} \geq 0. \tag{2.54}$$

By introducing the excess heat [86, 90],

$$\dot{S}^{\mathrm{ex}} \equiv \int \mathrm{d}x \; j \; \partial_x \ln p^{\mathrm{s}} \tag{2.55}$$

which has no definite sign, we can formulate the second law of steady-state thermodynamics [86, 87],

$$\dot{S}^{\mathrm{na}} = \dot{S} + \dot{S}^{\mathrm{ex}} \geq 0. \tag{2.56}$$

We have hence split the total entropy production into non-adiabatic and adiabatic contributions, which each correspond to a possible mechanism that lead to the breaking of time symmetry and hence to dissipation: the application of non-equilibrium constraints ($\dot{S}^{\mathrm{a}}$) and the presence of driving ($\dot{S}^{\mathrm{na}}$), and formulated a second law for the transition between steady states.

2.5.2 Multipartite systems

Our derivation from the previous system carries over to the case of two or more degrees of freedom without problems. A more interesting question is whether we can formulate inequalities like $\dot{S}^{\mathrm{na}} \geq 0$ on the level of individual subsystems, like we did for the total entropy production in Eq. (2.42).

Splitting the total entropy production of a subsystem like we did before, Eq. (2.51), we find

$$\dot{S}^{\mathrm{tot}} = \dot{S}^{\mathrm{na}} + \dot{S}^{\mathrm{a}} = \sum_n^N \dot{S}_n^{\mathrm{na}} + \sum_n^N \dot{S}_n^{\mathrm{a}} \tag{2.57}$$

where we define the non-adiabatic entropy production and the adiabatic entropy production due to the dynamics of X_n as

$$\dot{S}_n^{\mathrm{na}} \equiv \int \mathrm{d}\boldsymbol{x} \, \frac{p}{D_n} \left(\frac{j_n}{p} - \frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \geq 0 \tag{2.58}$$

and

$$\dot{S}_n^{\mathrm{a}} \equiv \int \mathrm{d}\boldsymbol{x}\ \frac{p}{D_n}\left(\frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}}\right)^2 \geq 0. \tag{2.59}$$

However, we note that on the level of a single subsystem,

$$\dot{S}_n^{\mathrm{tot}} = \dot{S}_n^{\mathrm{na}} + \dot{S}_n^{\mathrm{a}} + 2\int \mathrm{d}\boldsymbol{x}\ \frac{j_n^{\mathrm{s}}}{D_n}\frac{p}{p^{\mathrm{s}}}\left(\frac{j_n}{p} - \frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}}\right) \tag{2.60}$$

$$= \dot{S}_n^{\mathrm{na}} + \dot{S}_n^{\mathrm{a}} - 2\int \mathrm{d}\boldsymbol{x}\ (-\partial_n j_n^{\mathrm{s}})\frac{p}{p^{\mathrm{s}}} \tag{2.61}$$

where we used

$$\frac{j_n}{p} - \frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}} = -D_n\partial_n \ln\frac{p}{p^{\mathrm{s}}}. \tag{2.62}$$

Thus the splitting of the total entropy production into adiabatic and non-adiabatic contribution does not hold on the level of single subsystems. However, by summing (2.61) over all subsystems and using $\sum_n \partial_n j_n^{\mathrm{s}} = 0$, we obtain (2.57).

3 Building a model from data

We are now in a position to analyse the first learning task considered in this thesis: building a model for a set of data.

3.1 The learning problem

Imagine you are given the set of points $\left\{\boldsymbol{\xi}^1, \boldsymbol{\xi}^2, \ldots, \boldsymbol{\xi}^P\right\}$ shown in Fig. 3.1a for $\boldsymbol{\xi} \in \mathbb{R}^2$, with components drawn independently from the same distribution; here, a normal distribution with mean zero and standard deviation 1/2. These points $\boldsymbol{\xi}^\mu$ are then assigned a *true label* $\sigma_{\mathrm{T}}^\mu = \pm 1$ with equal probability. The labels are not correlated to the points or to the other labels. One possible such labelling is shown in Fig. 3.1b, where we are indicating points that have label $\sigma_{\mathrm{T}} = 1$ with blue triangles and those with label $\sigma_{\mathrm{T}} = -1$ with red triangles, respectively. While we are keeping the inputs fixed, different labellings are possible and another one is shown in Fig. 3.1c.

The learning task is to find a *model* that can predict these true labels for all the inputs $\boldsymbol{\xi}$. This model will be a function $\sigma(\boldsymbol{\xi}) = \pm 1$, parametrised by a number of

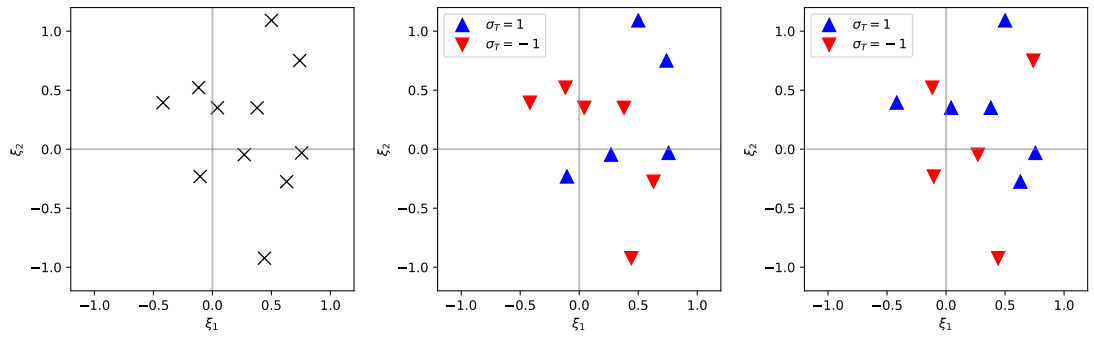


Figure 3.1 | Building a model from data. **a** A number of randomly drawn points in the $\mathbb{R}^2$ plane. **b** A possible assignment of true labels $\sigma_{\mathrm{T}} = \pm 1$, which were drawn at random and are not correlated to the points or to each other. **c** An alternative labelling of the same points.

parameters that we will call $\boldsymbol{w}$. For a given input set, the goal of learning is to find a functional form for σ and optimal parameters $\boldsymbol{w}$ such that the label given by the model, $\sigma(\boldsymbol{\xi})$, equals the true labels chosen at random $\sigma_{\mathrm{T}}(\boldsymbol{\xi})$ for as many inputs as possible.

In the following, we will first discuss a biological motivation for the kind of data that we will learn and for the neural network that we will use to learn the labelling. We will then apply the machinery developed in Chapter 2 to analyse the thermodynamic costs of building such a model for a set of input patterns with a given, true label. Our central result is that the thermodynamic costs of learning place a fundamental limit on the ability of the network to extract information from the data, which is ultimately based on the second law.

The results of this chapter have been published in Ref. [91].

3.2 Biological neural networks

Learning is about extracting models from sensory data. In living systems, it is implemented in neural networks, where a large number of nerve cells or neurons are connected to each other. These neurons communicate with each other using action potentials, the electric pulse used universally as the basic token of communication in neural systems [3]. Action potentials are transmitted between neurons via synapses, and their strength determines whether an incoming signal will make the receiving neuron trigger an action potential of its own that is then sent to other neurons.

We show a schematic snapshot of a simple neural network in Fig. 3.2. We are interested in the behaviour of one neuron in the network (red), which is connected to three input neurons (blue). The potential across the membranes of these neurons is shown in the inset graphs as a function of time. At one point in time, the input neurons 1 and 3 are simultaneously firing an action potential, *i.e.* they are sending an electrical signal to the red neuron. Input neuron 2 is silent. The neuron's task is to decide whether to fire an action potential or not given the input it receives from the three neurons that it is connected to. Each of the connections or synapses between the neurons has a weight, which indicates how strongly the receiving neuron is affected by the incoming signal. The red neuron decides whether to fire an action potential based on the weighted sum of the its inputs. Physiologically, the adaptation of the weights or synaptic strengths is a main mechanism for memory formation [3].

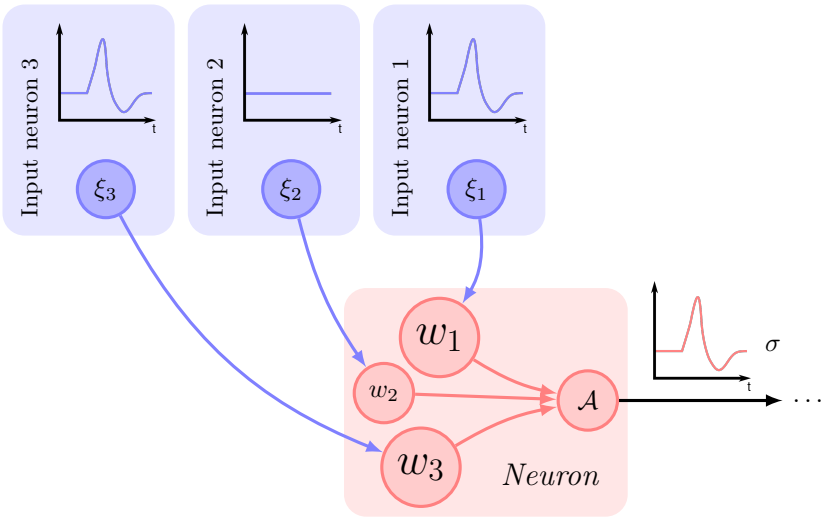


Figure 3.2 | Snapshot in time of a simple neural network. (a) A small neural network where the neuron of interest (red) is connected to three input neurons (blue), two of which are firing an action potential at that particular point in time while the second input neuron is silent. Each of the connections has a weight w_n. The neuron will also fire an action potential, depending on its activation $\mathcal{A}$, which is the weighted sum of its inputs. (b) We model the behaviour of the input neurons at a particular point in time by the input vectors $\boldsymbol{\xi} \in \{\pm 1\}^N$, since the precise temporal dynamics of the action potentials are not of interest for our purposes. Similarly, the response of the neuron is denoted by $\sigma = \pm 1$. In this example, $\boldsymbol{\xi} = \{1, -1, 1\}$ and $\sigma = 1$.

3.3 The perceptron and its dynamics

We model a neuron as a single-layer neural network or perceptron [8, 92], well known from machine learning and statistical physics*. The neuron makes N connections to other neurons and is fully characterised by the weights or synaptic strengths $\boldsymbol{w} \in \mathbb{R}^N$ of these connections, see figure 3.3. The neuron must learn whether it should fire an action potential or not for a set of P fixed input patterns or samples $\boldsymbol{\xi}^\mu = (\xi_1^\mu, \ldots, \xi_N^\mu)$, $\mu = 1, 2, \ldots, P$. Each input describes the activity of all the other connected neurons at a point in time. Since we are not interested in the precise temporal dynamics of the action potentials, we model the input of the neuron as vectors $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)$ where $\xi_n = 1$ if the nth connected neuron is firing an action potential in that input. For symmetry reasons, we set $\xi_n = -1$ if the nth neuron is silent. Every input has a fixed true label $\sigma_\mathrm{T}^\mu = \pm 1$, indicating whether an action potential should be fired in response to that input or not. These labels are independent of each other and equiprobable; once chosen, they remain

* Experimental justification for focusing on a single neuron comes from studies on psychophysical judgements in monkeys, which have been shown to depend on very few neurons [93].

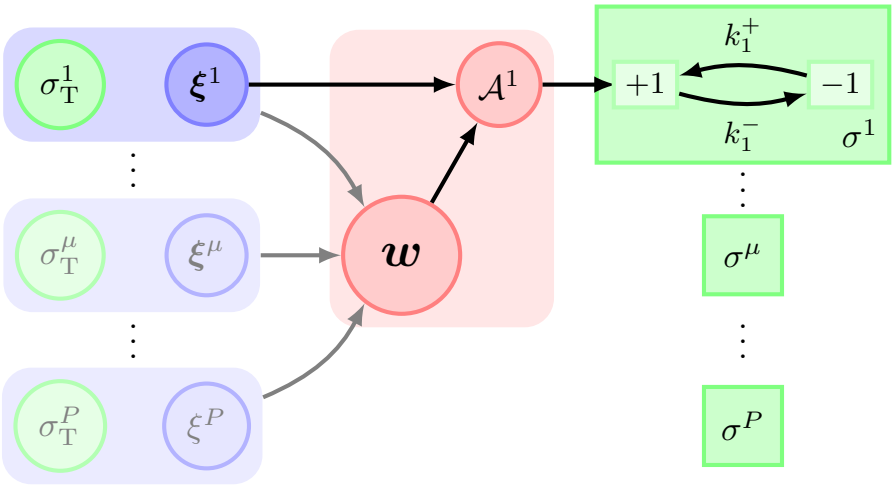


Figure 3.3 | Model of a single neuron. Given a set of inputs $\boldsymbol{\xi}^\mu \in \{\pm 1\}^N$ and their true labels $\sigma_{\mathrm{T}}^\mu = \pm 1$ (left), the neuron learns the mappings $\boldsymbol{\xi}^\mu \to \sigma_{\mathrm{T}}^\mu$ by adjusting its weights $\boldsymbol{w} \in \mathbb{R}^N$. It processes an input by computing the activation $\mathcal{A}^\mu = \boldsymbol{w} \cdot \boldsymbol{\xi}^\mu / \sqrt{N}$ which determines the transition rates of a two-state random process $\sigma^\mu = \pm 1$ indicating the label predicted by the neuron for each sample, shown here for $\mu = 1$.

fixed.

We model the label predicted by a neuron for each input $\boldsymbol{\xi}^\mu$ with a stochastic process $\sigma^\mu = \pm 1$ (right panel in figure 3.3). Assuming a thermal environment at fixed temperature T, the transition rates $k_\mu^\pm$ for these processes obey the detailed balance condition

$$k_\mu^+ / k_\mu^- = \exp\left(\mathcal{A}^\mu / k_{\mathrm{B}} T\right) \tag{3.1}$$

where k_{B} is Boltzmann's constant and $\mathcal{A}^\mu$ is the input-dependent activation

$$\mathcal{A}^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{w} \cdot \boldsymbol{\xi}^\mu \tag{3.2}$$

where the prefactor ensures the conventional normalisation. We interpret $p(\sigma^\mu = 1 \,|\, \boldsymbol{w})$ with fixed $\boldsymbol{\xi}^\mu$ as the probability that the μth input would trigger an action potential by the neuron. The goal of learning is to adjust the weights of the network $\boldsymbol{w}$ such that the predicted labels at any one time $\boldsymbol{\sigma} = \left(\sigma^1, \ldots, \sigma^P\right)$ equal the true labels $\boldsymbol{\sigma}_{\mathrm{T}} = \left(\sigma_{\mathrm{T}}^1, \ldots, \sigma_{\mathrm{T}}^P\right)$ for as many inputs as possible. A classic example for neurons performing the kind of associative learning described in the introduction are the Purkinje cells in the cerebellum [94, 95]. We have thus chosen the first half of the model, namely how we compute the label for a given input, which in this case is a stochastic function of the input and the weights $\boldsymbol{w}$.

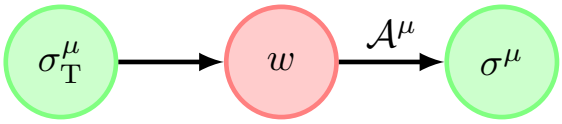


Figure 3.4 | Causal structure of the learning dynamics for a perceptron. The fixed labels $\sigma_{\mathrm{T}}^{\mu}$ exert a force on the weight w, which in turn determines the transition rates for the predicted labels σ^{μ} via the activations $\mathcal{A}^{\mu}$.

Let us describe the dynamics of learning by considering a network with a single weight learning one sample $\xi = \pm 1$ with label σ_{T}, *i.e.* $N = P = 1$. Here and throughout this chapter, we set $k_{\mathrm{B}} = T = 1$ to render energy and entropy dimensionless. The weight $w(t)$ obeys an overdamped Langevin equation [73]

$$\dot{w}(t) = -w(t) + f(w(t), \xi, \sigma_{\mathrm{T}}, t) + \zeta(t). \tag{3.3}$$

The total force on the weight arises from a harmonic potential $V(w) = w^2/2$, restricting the size of the weight*, and an external force $f(\cdot)$ introducing correlations between weight and input. The exact form of this "learning force" $f(\cdot)$ depends on the learning algorithm we choose. The thermal noise $\zeta(t)$ is Gaussian with correlations $\langle \zeta(t)\zeta(t') \rangle = 2\delta(t - t')$. Here and throughout, we use angled brackets to indicate averages over noise realisations, unless stated otherwise. We assume that initially at $t_0 = 0$, the weight is in thermal equilibrium, $p(w) \propto \exp(-w^2/2)$, and the labels are equiprobable, $p(\sigma_{\mathrm{T}}) = p(\sigma) = 1/2$. Choosing symmetric rates,

$$k^{\pm} = \gamma \exp(\pm \mathcal{A}/2), \tag{3.4}$$

the master equation [73] for the probability distribution $p(\sigma_{\mathrm{T}}, w, \sigma, t)$ with given ξ reads

$$\partial_t p(\sigma_{\mathrm{T}}, w, \sigma, t) = -\partial_w j_w(t) + j_\sigma(t), \tag{3.5}$$

where $\partial_t \equiv \partial/\partial t$ etc. and

$$j_w(t) = [-w + f(w, \xi, \sigma_{\mathrm{T}}, t) - \partial_w]\, p(\sigma_{\mathrm{T}}, w, \sigma, t), \tag{3.6a}$$
$$j_\sigma(t) = k^{\sigma} p(\sigma_{\mathrm{T}}, w, -\sigma, t) - k^{-\sigma} p(\sigma_{\mathrm{T}}, w, \sigma, t) \tag{3.6b}$$

are the probability currents for the weight and the predicted label, respectively. In splitting the total probability current for the system $(\sigma_{\mathrm{T}}, w, \sigma)$ into the currents (3.6), we have used the bipartite property of the system, *i.e.* that the thermal noise in each subsystem (w and σ), is independent of the other [38, 84]. We choose $\gamma \gg 1$, *i.e.* introduce a time-scale separation between the weights and the predicted labels, since a neuron processes a single input much faster than it learns.

* Restricting the size of the weights reflects experimental evidence suggesting the existence of an upper bound on synaptic strength in diverse nervous systems [96].

3.4 Efficiency of learning

The starting point to consider both the information-processing capabilities of the neuron and its non-equilibrium thermodynamics is the Shannon entropy of a random variable X with probability distribution $p(x)$,

$$S(X) \equiv -\sum_{x \in X} p(x) \ln p(x), \tag{3.7}$$

which is a measure of the uncertainty of X as discussed in Section 2.3. The natural quantity to measure the information learnt is the mutual information

$$I(\sigma_{\mathrm{T}} : \sigma) \equiv S(\sigma_{\mathrm{T}}) - S(\sigma_{\mathrm{T}}|\sigma) \tag{3.8}$$

which measures by how much, on average, the uncertainty about σ_{T} is reduced by knowing σ [31]. To discuss the efficiency of learning, we need to relate this information to the thermodynamic costs of adjusting the weight during learning from $t_0 = 0$ up to a time t, which are given by the well-known total entropy production [20] of the weight,

$$\Delta S_w^{\mathrm{tot}} \equiv \Delta S(w) + \Delta Q. \tag{3.9}$$

Here, ΔQ is the heat dissipated into the medium by the dynamics of the weight and $\Delta S(w)$ is the difference in Shannon entropy (3.7) of the marginalised distribution $p(w,t) = \sum_{\sigma_{\mathrm{T}},\sigma} p(\sigma_{\mathrm{T}}, w, \sigma, t)$ at times t_0 and t, respectively. We focus on the weights because they are the physical substrate of memory formation [3], while σ is just an auxiliary process without obvious physical equivalent in a biological network. We will show that in feedforward neural networks with Markovian dynamics (3.5, 3.6), the information learnt is bounded by the thermodynamic costs of learning,

$$I(\sigma_{\mathrm{T}} : \sigma) \leq \Delta S(w) + \Delta Q \tag{3.10}$$

for arbitrary learning algorithm $f(w, \xi, \sigma_{\mathrm{T}}, t)$ at all times $t > t_0$. This inequality is our first result.

We emphasise that while relations between changes in mutual information and total entropy production have appeared in the literature [36, 38, 39, 84, 85], they usually concern a single degree of freedom, say X, in contact with some other degree of freedom Y, and relate the change in mutual information $I(X : Y)$ due to the dynamics of X to the total entropy production of X. Instead, our relation connects the entropy production in the weights with the total change in mutual information between σ_{T} and σ, which is key for neural networks. Our derivation, see Sec. 3.A, builds on recent work by Horowitz [84] on information processing

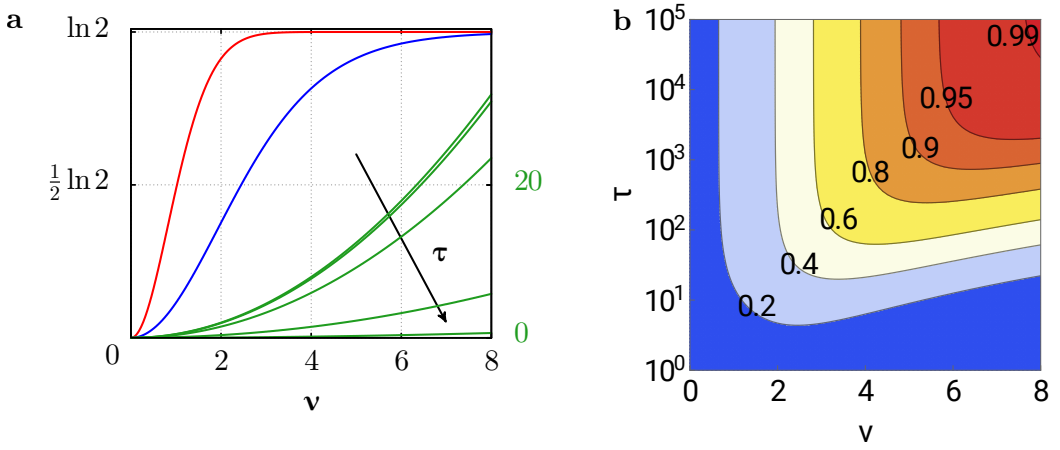


Figure 3.5 | Learning by a neuron with a single weight. b The final values for the entropy $S(w)$ of the weight (red), the mutual information $I(\sigma_{\mathrm{T}} : \sigma)$ (blue) are plotted as a function of the learning rate. They are state functions and hence do not depend on the learning process. On the other hand, the heat (green) is process dependent and plotted as a function of the learning rate ν for different process durations $\tau = 10^a$ where $a = -2, -1, 0, 1, 2$ from top to bottom. **b** We plot the efficiency η (3.11) for a neuron with a single weight learning a single sample as a function of the learning rate ν and learning duration τ in the limit $t \to \infty$.

in multipartite systems that we reviewed in Section 2.4. It can be generalised to N dimensions and P samples, see Eq. (3.16) below. Equation (3.10) suggests to introduce an efficiency of learning

$$\eta \equiv \frac{I(\sigma_{\mathrm{T}} : \sigma)}{\Delta S(w) + \Delta Q} \leq 1. \tag{3.11}$$

3.5 Toy model

As a first example, let us calculate the efficiency of Hebbian learning, a form of coincidence learning well known from biology [3, 97], for $N = P = 1$ in the limit $t \to \infty$. If the neuron should fire an action potential when its input neuron fires, or if they should both stay silent, *i.e.* $\xi = \sigma_{\mathrm{T}} = \pm 1$, the weight of their connection increases – "fire together, wire together". For symmetry reasons, the weight decreases if the input neuron is silent but the neuron should fire and vice versa, $\xi = -\sigma_{\mathrm{T}}$. This rule yields a final weight proportional to $\mathcal{F} \equiv \sigma_{\mathrm{T}}\xi$, so to minimise dissipation [98], we choose a learning force f linearly increasing with

time,

$$f(w, \xi, \sigma_{\mathrm{T}}, t) \equiv \begin{cases} \nu \mathcal{F} t/\tau & t \leq \tau \\ \nu \mathcal{F} & t > \tau, \end{cases} \tag{3.12}$$

where we have introduced the learning duration $\tau > 0$ and the factor $\nu > 0$ is conventionally referred to as the learning rate in the machine learning literature [92]. The total entropy production (3.9) can be computed from the distribution $p(\sigma_{\mathrm{T}}, w, t)$, which is obtained by first integrating σ out of equations (3.5, 3.6) and solving the resulting Fokker-Planck equation [75]. The total heat dissipated into the medium ΔQ is given by [20]

$$\begin{aligned} \Delta Q = \int_0^\infty \mathrm{d}t \int_{-\infty}^\infty \mathrm{d}w \; j_w(t) \left[-w(t) + f(w(t), \xi, \sigma_{\mathrm{T}}, t) \right] \\ = \frac{\nu^2 \mathcal{F}^2 (e^{-\tau} + \tau - 1)}{\tau^2}. \end{aligned} \tag{3.13}$$

As expected, no heat is dissipated in the limit of infinitely slow driving, $\lim_{\tau\to\infty} \Delta Q = 0$, while for a sudden potential switch $\tau \to 0$, $\lim_{\tau\to 0} \Delta Q = \nu^2 \mathcal{F}^2/2$. The change in Shannon entropy $\Delta S(w)$ is computed from the marginalised distribution $p(w, t) = \sum_{\sigma_{\mathrm{T}}} p(\sigma_{\mathrm{T}}, w, t)$. Finally, the mutual information (3.8) can be computed from the stationary solution of (3.5).

A plot of the efficiency (3.11), Fig. 3.5, highlights the two competing requirements for maximising η. First, all the information from the true label $S(\sigma_{\mathrm{T}}) = \ln 2$ needs to be stored in the weight by increasing the learning rate ν, which leads to $\Delta S(w) \to \ln 2$ and a strongly biased distribution $p(\sigma|w)$ such that $I(\sigma_{\mathrm{T}} : \sigma) \to \ln 2$. Second, we need to minimise the dissipated heat ΔQ, which increases with ν, by driving the weight slowly, $\tau \gg 1$.

3.6 More samples, higher dimensions

Moving on to a neuron with N weights $\boldsymbol{w}$ learning P samples with true labels $\boldsymbol{\sigma}_{\mathrm{T}} \equiv (\sigma_{\mathrm{T}}^1, \ldots, \sigma_{\mathrm{T}}^\mu, \ldots, \sigma_{\mathrm{T}}^P)$, we have a Langevin equation for each weight w_n with independent thermal noise sources $\zeta_n(t)$ such that $\langle \zeta_n(t) \zeta_m(t') \rangle = 2\delta_{nm}\delta(t - t')$ for $n, m = 1, \ldots, N$. Two learning scenarios are possible: *batch learning*, where the learning force is a function of all samples and their labels,

$$\dot{w}_n(t) = -w_n(t) + f(w_n(t), \{\xi_n^\mu, \sigma_{\mathrm{T}}^\mu\}, t) + \zeta_n(t). \tag{3.14}$$

A more realistic scenario from a biological perspective is *online learning*, where the learning force is a function of only one sample and its label at a time,

$$\dot{w}_n(t) = -w_n(t) + f(w_n(t), \xi_n^{\mu(t)}, \sigma_{\mathrm{T}}^{\mu(t)}, t) + \zeta_n(t). \tag{3.15}$$

The sample and label which enter this force are given by $\mu(t) \in \{1, \ldots, P\}$, which might be a deterministic function or a random process. Either way, the weights $\boldsymbol{w}$ determine the transition rates of the P independent two-state processes for the predicted labels $\boldsymbol{\sigma} \equiv (\sigma^1, \ldots, \sigma^\mu, \ldots, \sigma^P)$ via (3.1) and (3.2). Again, we assume that the thermal noise in each subsystem, w_n or σ^μ, is independent of all the others, and choose initial conditions at $t_0 = 0$ to be $p(\boldsymbol{w}) \propto \exp(-\boldsymbol{w} \cdot \boldsymbol{w}/2)$ and $p(\sigma^\mu_{\mathrm{T}}) = p(\sigma^\mu) = 1/2$. The natural quantity to measure the amount of learning after a time t in both scenarios is the sum of $I(\sigma^\mu_{\mathrm{T}} : \sigma^\mu)$ over all inputs. As we show in the Appendix 3.A, this information is bounded by the total entropy production of all the weights,

$$\sum_{\mu=1}^{P} I(\sigma^\mu_{\mathrm{T}} : \sigma^\mu) \leq \sum_{n=1}^{N} [\Delta S(w_n) + \Delta Q_n] = \sum_{n=1}^{N} \Delta S^{\mathrm{tot}}_n \tag{3.16}$$

where ΔQ_n is the heat dissipated into the medium by the nth weight and $\Delta S(w_n)$ is the change from t_0 to t in Shannon entropy (3.7) of the marginalised distribution $p(w_n, t)$. This is our main result.

Let us now compute the efficiency of online Hebbian learning in the limit $t \to \infty$. Since a typical neuron will connect to ~ 1000 other neurons [3], we take the thermodynamic limit by letting the number of samples P and the number of dimensions N both go to infinity while simultaneously keeping the ratio

$$\alpha \equiv P/N \tag{3.17}$$

on the order of one. The samples $\boldsymbol{\xi}^\mu$ are drawn at random from $p(\xi^\mu_n = 1) = p(\xi^\mu_n = -1) = 1/2$ and remain fixed*. We choose a learning force on the nth weight of the form (3.12) with $\mathcal{F} \to \mathcal{F}_n$ and assume that the process $\mu(t)$ is a random walk over the integers $1, \ldots, P$ changing on a timescale much shorter than the relaxation time of the weights. Since f^2 is finite, the learning force is effectively constant with

$$\mathcal{F}_n = \frac{1}{\sqrt{N}} \sum_{\mu=1}^{P} \xi^\mu_n \sigma^\mu_{\mathrm{T}}, \tag{3.18}$$

where the prefactor ensures the conventional normalisation [92]. Hence all the weights w_n are independent of each other and statistically equivalent. Averaging first over the noise with fixed $\boldsymbol{\sigma}_{\mathrm{T}}$, we find that w_n is normally distributed since the Langevin equation (3.15) defines an Ornstein–Uhlenbeck process w_n which for a Gaussian initial condition as we have chosen remains normally distributed [73]. Its

* In the limit of large N, only the first two moments of the distribution will matter, making this choice equivalent to sampling $\boldsymbol{\xi}^\mu$ from the surface of a hypersphere in N dimensions in that limit.

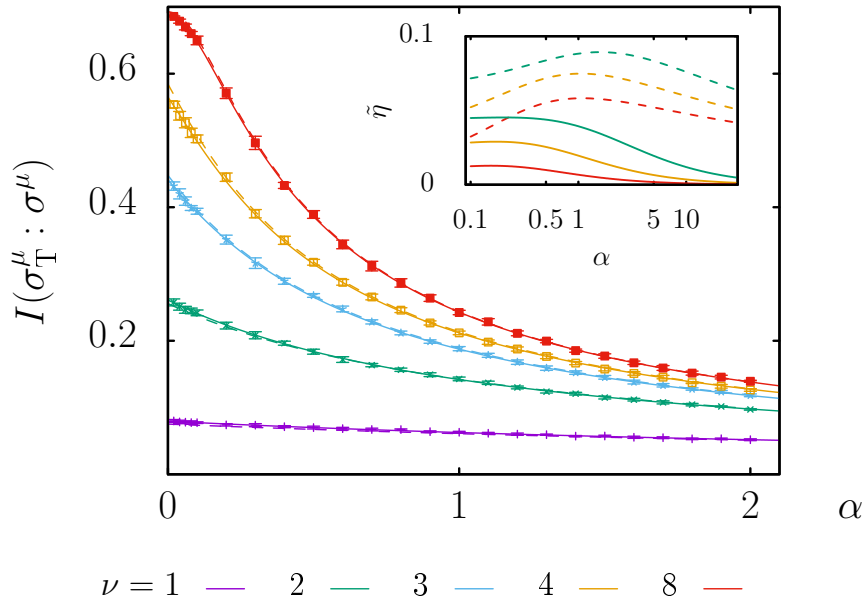


Figure 3.6 | Hebbian learning in the thermodynamic limit. We plot the mutual information between the true and predicted label of a randomly chosen sample (3.21) in the limit $t \to \infty$ with $N, P \to \infty$ as a function of $\alpha \equiv P/N$, computing p_{C}^{μ} from (3.20) (solid lines) and by Monte Carlo integration of $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$ (crosses, error bars indicate one standard deviation). The inset shows the learning efficiency (3.30) in the limits $\tau \to 0$ (solid) and $\tau \to \infty$ (dashed). In both plots, ν increases from bottom to top.

mean is $\langle w_n \rangle = \nu \mathcal{F}_n$ and its variance 1. The average with respect to the quenched disorder $\boldsymbol{\sigma}_{\mathrm{T}}$, which we shall indicate by an overline, is taken second by noting that $\mathcal{F}_n$ is normally distributed by the central limit theorem with $\overline{\mathcal{F}_n} = 0$ and $\overline{\mathcal{F}_n^2} = \alpha$, hence $\overline{\langle w_n \rangle} = 0$ and $\overline{\langle w_n^2 \rangle} = 1 + \alpha\nu^2$. The change in Shannon entropy of the marginalised distribution $p(w_n)$ is hence $\Delta S(w_n) = \ln(1 + \alpha\nu^2)$. Likewise, the heat dissipated by the nth weight $\overline{\Delta Q_n}$ is obtained by averaging Eq. (3.13) over $\mathcal{F} \to \mathcal{F}_n$.

The mutual information $I(\sigma_{\mathrm{T}}^{\mu} : \sigma^{\mu})$ is a functional of the marginalised distribution $p(\sigma_{\mathrm{T}}^{\mu}, \sigma^{\mu})$ which can be obtained by direct integration of $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$, see Appendix 3.B. Here we will take a simpler route starting from the *stability* of the μth sample [99]

$$\Delta^{\mu} \equiv \frac{1}{\sqrt{N}} \boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu} \sigma_{\mathrm{T}}^{\mu} = \mathcal{A}^{\mu} \sigma_{\mathrm{T}}^{\mu}. \tag{3.19}$$

Its role can be appreciated by considering the limit $T \to 0$, where it is easily verified using the detailed balance condition (3.1) that the neuron predicts the

correct label if and only if $\Delta^\mu > 0$. For $T = 1$, the neuron predicts the μth label correctly with probability

$$p_{\mathrm{C}}^\mu \equiv p(\sigma^\mu = \sigma_{\mathrm{T}}^\mu) = \int_{-\infty}^{\infty} \mathrm{d}\Delta^\mu \; p(\Delta^\mu) \frac{e^{\Delta^\mu}}{e^{\Delta^\mu} + 1} \tag{3.20}$$

where $p(\Delta^\mu)$ is the distribution generated by thermal noise and quenched disorder, yielding a Gaussian with mean ν and variance $1 + \alpha\nu^2$, see Appendix 3.B. The mutual information follows as

$$I(\sigma_{\mathrm{T}}^\mu : \sigma^\mu) = \ln 2 - S(p_{\mathrm{C}}^\mu) \tag{3.21}$$

with the shorthand for the entropy of a binary random variable $S(p) = -p \ln p - (1-p)\ln(1-p)$ [31]. It is plotted in Fig. 3.6 together with the mutual information obtained by Monte Carlo integration of $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$ with $N = 10000$. For a vanishing learning rate $\nu \to 0$ or infinitely many samples $\alpha \to \infty$, $p_{\mathrm{C}}^\mu \to 1/2$ and hence $I(\sigma_{\mathrm{T}}^\mu : \sigma^\mu) \to 0$. The maximum value $I(\sigma_{\mathrm{T}}^\mu : \sigma^\mu) = \ln 2$ is only reached for small α and decreases rapidly with increasing α, even for values of α where it is possible to construct a weight vector that classifies all the samples correctly [8]. This is a consequence of both the thermal noise in the system and the well-known failure of Hebbian learning to use the information in the samples perfectly [92].

While the integral in Eq. (3.20) has to be evaluated numerically, p_{C}^μ can be closely approximated analytically by $p(\Delta^\mu > 0)$ with the replacement $\nu \to \nu/2$. To that end, we first rewrite the sigmoid function in the integrand in terms of the hyperbolic tangent and exploit the similarity of the latter to the error function:

$$\begin{aligned} p_{\mathrm{C}}^\mu &= \int_{-\infty}^{\infty} \mathrm{d}\Delta^\mu \; p(\Delta^\mu) \frac{e^{\Delta^\mu/2}}{e^{\Delta^\mu/2} + e^{-\Delta^\mu/2}} && (3.22) \\ &= \frac{1}{2} + \frac{1}{2} \int_{-\infty}^{\infty} \mathrm{d}\Delta^\mu \; p(\Delta^\mu) \tanh(\Delta^\mu/2) && (3.23) \\ &\simeq \frac{1}{2} + \frac{1}{2} \int_{-\infty}^{\infty} \mathrm{d}\Delta^\mu \; p(\Delta^\mu) \operatorname{erf}(\gamma\Delta^\mu/2) && (3.24) \end{aligned}$$

where we choose $\gamma = 4/5$ by inspection of the graphs of the two functions. Now the convolution of a normal distribution and an error function has an exact solution,

$$\frac{1}{\sqrt{2\pi d^2}} \int_{-\infty}^{\infty} \mathrm{d}x \operatorname{erf}(ax + b) \exp\left(-\frac{(x-c)^2}{2d^2}\right) = \operatorname{erf}\left(\frac{b + ac}{\sqrt{1 + 2a^2d^2}}\right). \tag{3.25}$$

Setting $a = \gamma/2$, $b = 0$, $c = \nu$ and $d^2 = 1 + \alpha\nu^2$, we find that

$$p_{\mathrm{C}}^{\mu}(\alpha,\nu) \simeq \frac{1}{2} + \frac{1}{2}\operatorname{erf}\frac{\gamma\nu/2}{\sqrt{1+\gamma^2(1+\alpha\nu^2)/2}} \tag{3.26}$$

$$= \frac{1}{2} + \frac{1}{2}\operatorname{erf}\frac{\nu/2}{\sqrt{25/16+1/2+\alpha\nu^2/2}} \tag{3.27}$$

$$\simeq \frac{1}{2} + \frac{1}{2}\operatorname{erf}\frac{\nu/2}{\sqrt{2(1+\alpha\nu^2/4)}} \tag{3.28}$$

$$= p(\Delta^{\mu} > 0|\alpha,\nu/2), \tag{3.29}$$

where in the last line we recognise by inspection that our result is nothing but the integral over the distribution of stabilities $p(\Delta^{\mu}|\alpha,\nu/2)$ from 0 to ∞. The probability that the neuron predicts the correct label is hence given by the probability that the neuron learned the label correctly, $\Delta^{\mu} > 0$, with *half the learning rate*. We plot this expression with dashed lines in Fig. 3.6.

These results allow us to define the efficiency $\tilde{\eta}$ of Hebbian learning as a function of just α and ν,

$$\tilde{\eta} \equiv \alpha\frac{I(\sigma_{\mathrm{T}}^{\mu} : \sigma^{\mu})}{\Delta S(w_n) + \overline{\Delta Q_n}}, \tag{3.30}$$

where we have taken the mutual information per sample and the total entropy production per weight, multiplied by the number of samples and weights, respectively. Plotted in Fig. 3.6, this efficiency never reaches the optimal value 1, even in the limit of vanishing dissipation $\tau \to \infty$ (solid lines in Fig. 3.6).

3.7 Conclusion

We have introduced neural networks as models for studying the thermodynamic efficiency of building a model. This model has two parts. First, a description of the function that is used to compute the predicted label of a given input $\boldsymbol{\xi}^{\mu}$: compute the activation $\mathcal{A}^{\mu}$, Eq. (3.2), then draw the predicted label from the marginal distribution $p(\sigma^{\mu}|\boldsymbol{\xi}^{\mu})$. For the paradigmatic case of learning arbitrary binary labels for given inputs, we showed that the information acquired is bounded by the thermodynamic cost of learning. This is true for learning an arbitrary number of samples in an arbitrary number of dimensions for any learning algorithm without feedback for both batch and online learning.

Our framework opens up numerous avenues for further work. It will be interesting to analyse the efficiency of learning algorithms that employ feedback or use

an auxiliary memory [100]. Furthermore, we note that synaptic weight distributions are experimentally accessible [101, 102], offering the exciting possibility to test predictions on learning algorithms by looking at neural weight distributions. The inverse problem, *i.e.* deducing features of learning algorithms or the neural hardware that implements them by optimising some functional like the efficiency, looks like a formidable challenge, despite some encouraging progress in related fields [103, 104].

Appendices to chapter 3

The following appendices contain a detailed derivation of the main result of this chapter, inequality (3.16), in Sec. 3.A as well as additional analytical calculations for Hebbian learning in the thermodynamic limit in Sec. 3.B.

3.A Derivation of inequality (3.16)

We now give a detailed account of the dynamics of the neural networks and derive inequality (3.16) of the main text. To this end, we will use a lot of the machinery that we introduced in Sec. 2.4 to analyse the thermodynamics of the network. For simplicity, here we will focus on batch learning; the generalisation to online learning is straightforward. For a network with N weights $w_n \in \mathbb{R}$ learning P samples $\boldsymbol{\xi}^\mu \in \{\pm 1\}^N$ with their labels $\sigma_\mathrm{T}^\mu = \pm 1$, $\mu = 1, 2, \ldots, P$, we have the N Langevin equations (3.14), which we repeat here for convenience

$$\dot{w}_n(t) = -w_n(t) + f(w_n(t), \{\xi_n^\mu, \sigma_\mathrm{T}^\mu\}, t) + \zeta_n(t). \tag{3.31}$$

The Gaussian noise $\zeta_n(t)$ has correlations $\langle \zeta_n(t)\zeta_m(t')\rangle = 2T\delta_{nm}\delta(t-t')$ for $n, m = 1, \ldots, N$ where T is the temperature of the surrounding medium and we have set Boltzmann's constant to unity to render entropy dimensionless. The weights $\boldsymbol{w}$ determine the transition rates of the P independent two-state processes for the predicted labels σ^μ via

$$k_\mu^+ / k_\mu^- = \exp\left(\mathcal{A}^\mu / T\right) \tag{3.32}$$

where $\mathcal{A}^\mu$ is the input-dependent activation

$$\mathcal{A}^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{w} \cdot \boldsymbol{\xi}^\mu \tag{3.33}$$

For the remainder of this section, we set $T = 1$, rendering energy dimensionless. We assume that the thermal noise in each subsystem, like w_n or σ^μ, is independent of all the others. This multipartite assumption [84] allows us to write the master equation for the distribution $p(\boldsymbol{\sigma}_\mathrm{T}, \boldsymbol{w}, \boldsymbol{\sigma}, t)$ with $\boldsymbol{\sigma}_\mathrm{T} \equiv (\sigma_\mathrm{T}^1, \ldots, \sigma_\mathrm{T}^P)$ and $\boldsymbol{\sigma} \equiv (\sigma^1, \ldots, \sigma^P)$ as

$$\partial_t p(\boldsymbol{\sigma}_\mathrm{T}, \boldsymbol{w}, \boldsymbol{\sigma}, t) = -\sum_{n=1}^{N} \partial_n j_n(t) + \sum_{\mu=1}^{P} j_\mu(t), \tag{3.34}$$

where $\partial_t \equiv \partial/\partial t$, $\partial_n \equiv \partial/\partial\, w_n$ and the probability currents for the nth weight w_n and the μth predicted label σ^μ are given by

$$j_n(t) = \Big[-w_n + f(w_n, \boldsymbol{\xi}^{\mu(t)}, \sigma_{\mathrm{T}}^{\mu(t)}, t) - \partial_n\Big]\, p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma}, t), \tag{3.35a}$$

$$j_\mu(t) = k^+ p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \sigma^1, \ldots, -\sigma^\mu, \ldots, \sigma^P, t) - k^- p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma}, t). \tag{3.35b}$$

We choose symmetric rates $k_\mu^\pm = \gamma \exp(\pm \mathcal{A}^\mu/2)$ with $\gamma \gg 1$. Initially, the true labels $\boldsymbol{\sigma}_{\mathrm{T}}$, weights $\boldsymbol{w}$ and predicted labels are all uncorrelated with

$$p_0(\sigma_{\mathrm{T}}^\mu) = 1/2, \tag{3.36}$$

$$p_0(\sigma^\mu) = 1/2, \quad \text{and} \tag{3.37}$$

$$p_0(\boldsymbol{w}) = \frac{1}{(2\pi)^{N/2}} \exp(-\boldsymbol{w}\cdot\boldsymbol{w}/2). \tag{3.38}$$

Since the following discussion applies to the time-dependent dynamics (3.34), we understand that all quantities that will be introduced in the remainder of this section have an implicit time-dependence via the distribution $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma}, t)$ or the currents (3.35).

We start our derivation of the main inequality, equation (3.16) of the main text, from the refined second law for the nth weight (see Sec. 2.4 for a detailed discussion)

$$\partial_t S(w_n) + \dot{Q}_n - l_n(w_n; \boldsymbol{\sigma}_{\mathrm{T}}) \geq 0, \tag{3.39}$$

with equality in equilibrium only. Here,

$$\dot{Q}_n = \sum_{\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{w}\; j_n(t) F_n(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma}) \tag{3.40}$$

is the heat dissipated by the nth weight into its isothermal bath which experiences a total force $F_n = -w_n(t) + f(w_n(t), \{\xi_n^\mu, \sigma_{\mathrm{T}}^\mu\}, t)$. We write the Shannon entropy of the entire system as $S(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$ in a slight abuse of notation to emphasise that we consider the Shannon entropy of the full distribution $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$. The change of the Shannon entropy of the entire system contributes two terms to the second law (3.39): First, we have the change of Shannon entropy of the marginalised distribution $p(w_n)$,

$$\dot{S}_n(w_n) = -\sum_{\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{\sigma}} \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{w}\; j_n(t) \partial_n \ln p(w_n) = \partial_t S(w_n), \tag{3.41}$$

The second contribution is the refined thermodynamic learning rate (2.47), which is a thermodynamically consistent measure of how much the dynamics of w_n change

the mutual information $I(w_n : \boldsymbol{\sigma}_\mathrm{T}, \overline{\boldsymbol{w}}, \boldsymbol{\sigma})$, in particular for a system that continuously rewrites a single memory [105]. We obtain it from the full thermodynamic learning rate (2.41) by exploiting the causal structure of the dynamics, as we discussed in Sec. 2.4. Hence, the refined thermodynamic learning rate includes only the interactions of the nth weight with the degrees of freedom that it directly interacts with, *i.e.* the true labels $\boldsymbol{\sigma}_\mathrm{T}$,

$$l_n(w_n; \boldsymbol{\sigma}_\mathrm{T}) = \sum_{\boldsymbol{\sigma}, \boldsymbol{\sigma}_\mathrm{T}} \int_{-\infty}^{\infty} \mathrm{d}\boldsymbol{w}\; j_n(t) \partial_n \ln p(\boldsymbol{\sigma}_\mathrm{T} | w_n). \tag{3.42}$$

The refined second law of stochastic thermodynamics for neural networks yields N inequalities of the form (3.39). Integrating over time and summing over all the weights, we find

$$\sum_{n=1}^{N} [\Delta S(w_n) + \Delta Q_n] \geq \sum_{n=1}^{N} \int_0^\infty \mathrm{d}t\; l_n(w_n; \boldsymbol{\sigma}_\mathrm{T}) = \sum_{n=1}^{N} \Delta I(w_n : \boldsymbol{\sigma}_\mathrm{T}) \tag{3.43}$$

The precise definition of all the terms were discussed previously in Section 3.6. The crucial point for the last equality is that the labels $\boldsymbol{\sigma}_\mathrm{T}$ are static, so that the mutual information $I(w_n : \boldsymbol{\sigma}_\mathrm{T})$ changes only due to the dynamics of w_n and hence $\partial_t I(w_n : \boldsymbol{\sigma}_\mathrm{T}) = l_n(w_n; \boldsymbol{\sigma}_\mathrm{T})$. If we restricted ourselves to online learning, where the learning force is a local force with only one sample and its label acting on the weights, we could consider this as an upper bound on the amount of information that the weights can acquire during learning, yielding the same result for the efficiency.

To make progress towards our main result, we need to show that

$$\sum_{n=1}^{N} \Delta I(w_n : \boldsymbol{\sigma}_\mathrm{T}) \geq \sum_{\mu=1}^{P} \Delta I(\sigma_\mathrm{T}^\mu : \sigma^\mu). \tag{3.44}$$

First, we note that from the chain rule of mutual information [31], we have

$$\Delta I(\boldsymbol{w} : \boldsymbol{\sigma}_\mathrm{T}) = \Delta I(w_1, \ldots, w_n : \boldsymbol{\sigma}_\mathrm{T}) = \sum_{n=1}^{N} \Delta I(w_n : \boldsymbol{\sigma}_\mathrm{T} \,|\, w_{n-1}, \ldots, w_1) \tag{3.45}$$

with the conditional mutual information [31]

$$I(w_n : \boldsymbol{\sigma}_\mathrm{T} \,|\, w_{n-1}, \ldots, w_1) \equiv S(w_n | w_{n-1}, \ldots, w_1) - S(w_n | \boldsymbol{\sigma}_\mathrm{T}, w_{n-1}, \ldots, w_1). \tag{3.46}$$

Due to the form of the Langevin equation for the single weight, Eq. (3.31), individual weights are uncorrelated, and hence the conditional mutual information

simplifies to

$$\Delta I(w_n : \boldsymbol{\sigma}_\mathrm{T} \,|\, w_{n-1}, \ldots, w_1) = \Delta S(w_n|w_{n-1}, \ldots, w_1) - \Delta S(w_n|\boldsymbol{\sigma}_\mathrm{T}, w_{n-1}, \ldots, w_1) \tag{3.47}$$

$$= \Delta S(w_n) - \Delta S(w_n|\boldsymbol{\sigma}_\mathrm{T}) \tag{3.48}$$

$$= \Delta I(w_n : \boldsymbol{\sigma}_\mathrm{T}) \tag{3.49}$$

such that

$$\sum_{n=1}^{N} \Delta I(w_n : \boldsymbol{\sigma}_\mathrm{T}) = \Delta I(\boldsymbol{w} : \boldsymbol{\sigma}_\mathrm{T}). \tag{3.50}$$

Next, we show that

$$\Delta I(\boldsymbol{w} : \boldsymbol{\sigma}_\mathrm{T}) = \sum_{\mu=1}^{P} \Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu \,|\, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) \overset{!}{\geq} \sum_{\mu=1}^{P} \Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu). \tag{3.51}$$

using the independence of the given labels $\boldsymbol{\sigma}_\mathrm{T}$. We first note that

$$\Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu \,|\, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) = \Delta S(\sigma_\mathrm{T}^\mu|\sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) - \Delta S(\sigma_\mathrm{T}^\mu|\boldsymbol{w}, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) \tag{3.52}$$

$$= \Delta S(\sigma_\mathrm{T}^\mu) - \Delta S(\sigma_\mathrm{T}^\mu|\boldsymbol{w}, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) \tag{3.53}$$

while

$$\Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu) = \Delta S(\sigma_\mathrm{T}^\mu) - \Delta S(\sigma_\mathrm{T}^\mu|\boldsymbol{w}) \tag{3.54}$$

Hence for $\Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu \,|\, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) \overset{!}{\geq} \Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu)$, we need

$$\Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu \,|\, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) - \Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu) \tag{3.55}$$

$$= \Delta S(\sigma_\mathrm{T}^\mu|\boldsymbol{w}) - \Delta S(\sigma_\mathrm{T}^\mu|\boldsymbol{w}, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) \tag{3.56}$$

$$= \Delta I(\sigma_\mathrm{T}^\mu : \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1 \,|\, \boldsymbol{w}) \tag{3.57}$$

$$\geq 0 \tag{3.58}$$

where we first used that the σ_T^μ are independent and identically distributed. The last inequality follows since any mutual information, conditional or not, is always greater than or equal to zero [31]. We have thus shown that $\Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu \,|\, \sigma_\mathrm{T}^{\mu-1}, \ldots, \sigma_\mathrm{T}^1) \geq \Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu)$ and hence (3.51) is true.

Finally, to prove that $\Delta I(\boldsymbol{w} : \sigma_\mathrm{T}^\mu) > \Delta I(\sigma_\mathrm{T}^\mu : \sigma^\mu)$, we consider the full probability distribution $p(\boldsymbol{\sigma}_\mathrm{T}, \boldsymbol{w}, \boldsymbol{\sigma})$. From the master equation, Eq. (3.34), we can write this distribution as

$$p(\boldsymbol{\sigma}_\mathrm{T}, \boldsymbol{w}, \boldsymbol{\sigma}) = p(\boldsymbol{\sigma}_\mathrm{T}) p(\boldsymbol{w}|\boldsymbol{\sigma}_\mathrm{T}) \left[p^{(0)}(\boldsymbol{\sigma}|\boldsymbol{w}) + \frac{1}{\gamma} p^{(1)}(\boldsymbol{\sigma}|\boldsymbol{w}) + \mathcal{O}(1/\gamma^2) \right] \tag{3.59}$$

with $\gamma \gg 1$ for physiological reasons as described in the text – it takes the neuron longer to learn than to generate an action potential. Hence to first order, $\boldsymbol{\sigma}_{\mathrm{T}} \to \boldsymbol{w} \to \boldsymbol{\sigma}$ is by definition a Markov chain [31]. Integrating out all the labels, true and predicted, except for the μth one, we have the Markov chain $\sigma_{\mathrm{T}}^{\mu} \to \boldsymbol{w} \to \sigma^{\mu}$. For such a Markov chain, it is easy to show the following data processing inequality [31],

$$\Delta I(\sigma_{\mathrm{T}}^{\mu} : \boldsymbol{w}) \geq \Delta I(\sigma_{\mathrm{T}}^{\mu} : \sigma^{\mu}), \tag{3.60}$$

which completes our derivation.

3.B Hebbian learning in the thermodynamic limit

In this section, we return to Hebbian learning in the thermodynamic limit for long times $t \to \infty$ and give detailed calculations for the distribution $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$ and $p(\Delta^{\mu})$ and derive an analytical approximation for the mutual information $I(\sigma_{\mathrm{T}} : \sigma)$.

3.B.1 Direct integration of the full distribution $p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma})$

To compute the mutual information between the true and predicted label of a given sample, $I(\sigma_{\mathrm{T}}^{\mu} : \sigma^{\mu})$, we need the distribution $p(\sigma_{\mathrm{T}}^{\mu}, \sigma^{\mu})$ or, since both $\sigma_{\mathrm{T}}^{\mu}$ and σ^{μ} are symmetric binary random variables, the probability that $\sigma_{\mathrm{T}}^{\mu} = \sigma^{\mu}$. Our aim in this section is to obtain this probability for Hebbian learning in the thermodynamic limit with $t \to \infty$ by direct integration of the full distribution over the true labels, weights and predicted labels for a given set of samples $\{\boldsymbol{\xi}^{\mu}\}$, which will also give additional motivation for introducing the stability Δ^{μ} of a sample.

We start with the full probability distribution

$$p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \boldsymbol{\sigma}) = \left(\frac{1}{2}\right)^{P} \left(\prod_{n=1}^{N} \frac{e^{-(w_n - \nu \mathcal{F}_n)^2/2}}{\sqrt{2\pi}} \right) \left(\prod_{\mu=1}^{P} \frac{e^{\sigma^{\mu} \boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu}/2\sqrt{N}}}{e^{-\boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu}/2\sqrt{N}} + e^{\boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu}/2\sqrt{N}}} \right), \tag{3.61}$$

where ν is the learning rate and $\mathcal{F}_n$ is a suitably scaled average over the samples and labels,

$$\mathcal{F}_n = \frac{1}{\sqrt{N}} \sum_{\rho=1}^{P} \sigma_{\mathrm{T}}^{\rho} \xi_n^{\rho} \tag{3.62}$$

While the sum over the predicted labels $\sigma^{\rho \neq \mu} = \pm 1$ is trivial, we can integrate over the true labels by noting that we can rewrite the exponent as

$$p(\boldsymbol{\sigma}_{\mathrm{T}}, \boldsymbol{w}, \sigma^{\mu}) = \left(\frac{1}{2}\right)^{P} \left(\prod_{n=1}^{N} \frac{e^{-(w_n - \nu \sigma_{\mathrm{T}}^{\mu} \xi_n^{\mu}/\sqrt{N} - \nu \mathcal{F}_n^{\overline{\mu}})^2/2}}{\sqrt{2\pi}} \right) \frac{e^{\sigma^{\mu} \boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu}/2\sqrt{N}}}{e^{-\boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu}/2\sqrt{N}} + e^{\boldsymbol{w} \cdot \boldsymbol{\xi}^{\mu}/2\sqrt{N}}}$$

(3.63)

where the only dependence of the weight distribution on the true labels $\sigma_{\mathrm{T}}^{\rho\neq\mu}$ is now confined to the sum

$$\mathcal{F}_n^{\overline{\mu}} \equiv \frac{1}{\sqrt{N}} \sum_{\rho\neq\mu}^{P} \sigma_{\mathrm{T}}^{\rho} \xi_n^{\rho}. \tag{3.64}$$

In the thermodynamic limit, this allows us to replace the sum over all $\sigma_{\mathrm{T}}^{\mu\neq\rho}$ by an integral over the stochastic variable $\mathcal{F}_n^{\overline{\mu}}$, which is normally distributed by the central limit theorem and has mean 0 and variance α. Carrying out the integral, we find

$$p(\sigma_{\mathrm{T}}^{\mu}, \boldsymbol{w}, \sigma^{\mu}) = \left(\prod_{n=1}^{N} \frac{e^{-(w_n - \nu\sigma_{\mathrm{T}}^{\mu}\xi_n^{\mu}/\sqrt{N})^2/2(1+\alpha\nu^2)}}{\sqrt{2\pi(1+\alpha\nu^2)}} \right) \frac{e^{\sigma^{\mu}\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}/2\sqrt{N}}}{e^{-\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}/2\sqrt{N}} + e^{\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}/2\sqrt{N}}} \tag{3.65}$$

Since both $\sigma_{\mathrm{T}}^{\mu}$ and σ^{μ} are binary random variables and $\sigma_{\mathrm{T}}^{\mu} = \pm 1$ with equal probabilities, the mutual information between the true and predicted label can be written as

$$I(\sigma_{\mathrm{T}}^{\mu} : \sigma^{\mu}) = \ln 2 - S[p(\sigma_{\mathrm{T}}^{\mu} = \sigma^{\mu})] \tag{3.66}$$

with the shorthand for the binary entropy $S[p] = -p\ln p - (1-p)\ln(1-p)$ [31]. With $\sigma^{\mu} = \sigma_{\mathrm{T}}^{\mu}$ in the exponential term of Eq. (3.65) and noting that $(\sigma_{\mathrm{T}}^{\mu}\xi_n^{\mu})^2 = 1$ for all $\sigma_{\mathrm{T}}^{\mu}$, ξ_n^{μ}, we then have

$$p(\sigma_{\mathrm{T}}^{\mu} = \sigma^{\mu}, \boldsymbol{w}) = \left(\prod_{n=1}^{N} \frac{e^{-(w_n\sigma_{\mathrm{T}}^{\mu}\xi_n^{\mu} - \nu/\sqrt{N})^2/2(1+\alpha\nu^2)}}{\sqrt{2\pi(1+\alpha\nu^2)}} \right) \frac{e^{\sigma_{\mathrm{T}}^{\mu}\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}/2\sqrt{N}}}{e^{-\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}/2\sqrt{N}} + e^{\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}/2\sqrt{N}}} \tag{3.67}$$

It thus becomes clear that $\boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}\sigma_{\mathrm{T}}^{\mu}$ is the sum of N random variables with mean $\nu/\sqrt{N}$ and variance $1+\alpha\nu^2$. We are then motivated to introduce the stability of a sample,

$$\Delta^{\mu} \equiv \frac{1}{\sqrt{N}} \boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu}\sigma_{\mathrm{T}}^{\mu} = \mathcal{A}^{\mu}\sigma_{\mathrm{T}}^{\mu}. \tag{3.68}$$

which, from Eq. (3.67), is normally distributed with mean ν and variance $1+\alpha\nu^2$. Introducing the stability allows us to replace the integral over all the weights by an integral over the stability,

$$p(\sigma_{\mathrm{T}}^{\mu} = \sigma^{\mu}) = \int_{-\infty}^{\infty} \mathrm{d}\Delta^{\mu} \frac{e^{-(\Delta^{\mu}-\nu)^2/2(1+\alpha\nu^2)}}{\sqrt{2\pi(1+\alpha\nu^2)}} \frac{e^{\Delta^{\mu}}}{1+e^{\Delta^{\mu}}} = \int_{-\infty}^{\infty} \mathrm{d}\Delta^{\mu} p(\Delta^{\mu}) \frac{e^{\Delta^{\mu}}}{1+e^{\Delta^{\mu}}} \tag{3.69}$$

which is the distribution obtained as Eq. (3.20) of the main text.

3.B.2 Direct derivation of the distribution of stabilities

Let us quickly show how the distribution of stabilities

$$\Delta^\mu \equiv \frac{1}{\sqrt{N}} \boldsymbol{w} \cdot \boldsymbol{\xi}^\mu \sigma_{\mathrm{T}}^\mu, \tag{3.70}$$

$\mu = 1, \ldots, P$, is obtained directly from its definition. The weights are given by

$$\boldsymbol{w} = \frac{1}{\sqrt{N}} \nu \sum_{\rho=1}^{P} \boldsymbol{\xi}[\rho] \sigma_{\mathrm{T}}^\rho + \boldsymbol{y} \tag{3.71}$$

with $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$ where y_n are normally distributed random variables with mean 0 and variance 1 arising from the thermal fluctuations in equilibrium. Substituting Eq. (3.71) into Eq. (3.70), we have

$$\begin{aligned} \Delta^\mu =& \frac{1}{N} \nu \sum_{\rho=1}^{P} \sigma_{\mathrm{T}}^\rho \sigma_{\mathrm{T}}^\mu \boldsymbol{\xi}[\rho] \cdot \boldsymbol{\xi}^\mu + \frac{1}{\sqrt{N}} \sigma_{\mathrm{T}}^\mu \boldsymbol{\xi}^\mu \cdot \boldsymbol{y} & (3.72) \\ =& \nu + \frac{1}{N} \nu \sum_{\rho \neq \mu}^{P} \sigma_{\mathrm{T}}^\rho \sigma_{\mathrm{T}}^\mu \boldsymbol{\xi}[\rho] \cdot \boldsymbol{\xi}^\mu + \frac{1}{\sqrt{N}} \sigma_{\mathrm{T}}^\mu \boldsymbol{\xi}^\mu \cdot \boldsymbol{y} & (3.73) \end{aligned}$$

where going to the last line we have used the fact that $\boldsymbol{\xi}^\mu \cdot \boldsymbol{\xi}^\mu = N$. By inspection, we see that the second term is the sum of $N(P-1) \approx NP$ random numbers $\pm\nu/N$ and the last term is the sum of N random numbers $y_n/\sqrt{N}$. By the central limit theorem, Δ^μ is hence normally distributed with mean $\overline{\langle \Delta^\mu \rangle} = \nu$ and variance

$$\overline{\langle (\Delta^\mu)^2 \rangle} - \overline{\langle \Delta^\mu \rangle}^2 = \nu^2 + NP \frac{\nu^2}{N^2} + N \frac{1}{N} - \nu^2 = 1 + \alpha \nu^2. \tag{3.74}$$

4 Generalising from examples

We continue to study the perceptron, parametrised by its set of weights $\boldsymbol{w}$, which assigns labels $\sigma^\mu = \pm 1$ to inputs $\boldsymbol{\xi}^\mu \in \mathbb{R}^N$. In the previous chapter, the aim of the perceptron was to reproduce as faithfully as possible the true labels, drawn at random, for a fixed set of inputs, see Fig. 3.1. Here, we will study a different learning problem where the true labels are generated according to some fixed function $\sigma_\mathrm{T}(\xi) = \pm 1$, which is parametrised by a set of parameters $\boldsymbol{T}$. These could for example be the weights of another neural network, which is often called the teacher [8, 92].

The perceptron, called the student in this context, has to infer this rule from a number of examples $(\boldsymbol{\xi}, \sigma_\mathrm{T})$ supplied by the teacher. Our focus in this chapter is on the final step of information processing: how well can the network emulate the function implemented by the teacher after a training period, *i.e.* how well do the outputs of the student, σ, match the correct output of the teacher σ_T for the any given input? We will show that the ability of the network to generalise such a rule from the examples it has seen to previously unseen inputs is bound by the dissipation of free energy by the components of the network as a consequence of the second law of stochastic thermodynamics.

Our results apply to a wide variety of learning algorithms. For illustration purposes, we analyse three learning algorithms in particular: Hebbian learning [97, 106], which was inspired by the neurobiology of memory formation; the celebrated Perceptron algorithm [107], whose discovery led to a surge in interest in neural networks in the 1960s and which is still very influential; and finally AdaTron learning [108], a refinement of the Perceptron algorithm with surprising dynamical features.

This chapter is organised as follows. We give a detailed description of our model and its dynamics in Sec. 4.1 and 4.2. We derive a general bound in Sec. 4.3 and discuss a number of simple examples with different learning algorithms in Sec. 4.4. We then derive a second, sharper bound in Sec. 4.5 and analyse the efficiency of learning in large networks in Sec. 4.6. We give some concluding perspectives in Sec. 4.7. Detailed proofs and a number of technical points are discussed in the appendices. The results from this chapter have been published in [109].

4.1 Inputs and labels, Teacher and student

We consider a single neuron, modeled by a perceptron as discussed in Chapter 3. The inputs are distributed according to

$$p(\boldsymbol{\xi}) = \prod_{n=1}^{N} \frac{1}{2} \left[\delta(\xi_n - 1) + \delta(\xi_n + 1)\right]. \tag{4.1}$$

The neuron itself is fully characterised by the N weights $\boldsymbol{w} \in \mathbb{R}^N$ of its N afferent connections. The weights obey noisy dynamics, to be specified in Sec. 4.2. Presented with a given input $\boldsymbol{\xi}$, the neuron computes an input-dependent activation

$$\mathcal{A} \equiv \frac{1}{\sqrt{N}} \boldsymbol{w} \cdot \boldsymbol{\xi} \tag{4.2}$$

where the prefactor ensures normalisation. The activation determines whether the neuron will fire an action potential or not, $\sigma = 1$ or -1, respectively. If the prediction was noise-free, we would have $\sigma = \mathrm{sgn}(\mathcal{A})$, where $\mathrm{sgn}(x > 0) = 1$ and $\mathrm{sgn}(x \leq 0) = -1$; instead, the predicted label σ is stochastic with

$$p(\sigma|\mathcal{A}) \propto \exp(\beta \sigma \mathcal{A}) \tag{4.3}$$

where β is the inverse temperature of the surrounding heat bath. As we have done previously, we set $k_\mathrm{B} = \beta = 1$ for the remainder of this chapter, rendering entropy and energy dimensionless without loss of generality.

The rules we want to learn are Boolean functions of the inputs. More precisely, we will focus on realisable rules which are linearly separable, *i.e.* we can write

$$\sigma_\mathrm{T} = \mathrm{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi}) = \pm 1 \tag{4.4}$$

where the teacher network $\boldsymbol{T} \in \mathbb{R}^N$ has the same architecture as the neural network $\boldsymbol{w}$. This function is shown schematically in Fig. 4.1 for two dimensions, where the teacher vector separates the plane from which the inputs are drawn into two halves. In N dimensions, the function (4.4) separates the N-dimensional hypersphere from which the inputs $\boldsymbol{\xi}$ are drawn into two hemispheres, one with $\sigma_\mathrm{T} = 1$ and one with $\sigma_\mathrm{T} = -1$, with the vector $\boldsymbol{T}$ pointing to the "northpole".

The components of the teacher are independent and drawn from a normal distribution with mean 0 and variance 1 and kept fixed. We draw the teacher at random in order to make general statements about the ability of the network to infer teachers of this form. By analogy, the neuron in such a setup is often called the student. We can interpret the true label of an input as an indication of whether the student should fire an action potential in response to that input or not. We

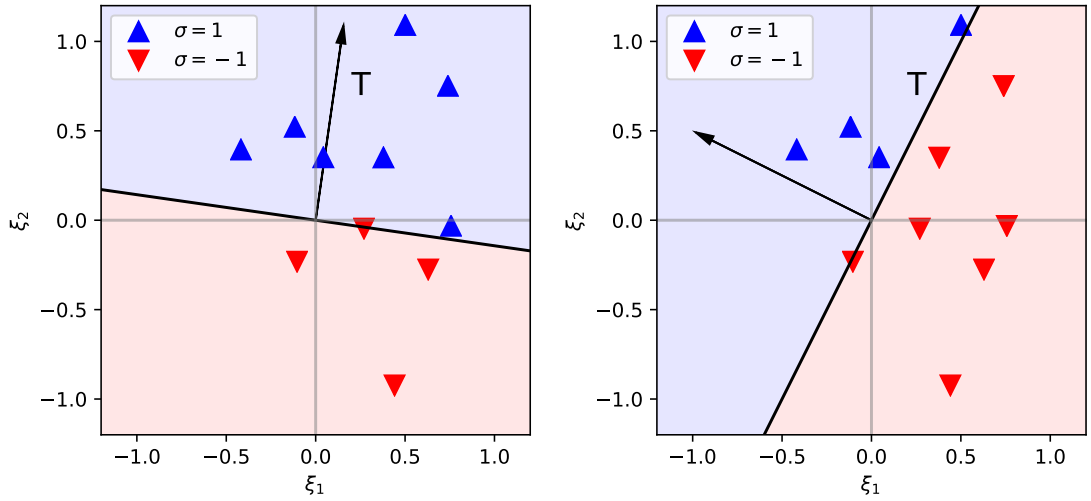


Figure 4.1 | Learning from a teacher. For inputs $\boldsymbol{\xi} \in \mathbb{R}^2$, we plot the functions implemented by two neural networks with different weight vectors $\boldsymbol{T} \in \mathbb{R}^2$ implementing the rule $\sigma_\mathrm{T} = \mathrm{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi})$. All the points in the regions shaded blue will be assigned $\sigma_\mathrm{T} = 1$, while points in the region shaded red will have a true label $\sigma_\mathrm{T} = -1$. The aim of the student is to infer the parameters $\boldsymbol{T}$ from a number of samples $(\sigma_\mathrm{T}, \boldsymbol{\xi})$ for randomly drawn inputs $\boldsymbol{\xi}$. N.B. inputs in this plot are drawn from a normal distribution with mean 0 and variance 0.5 rather than the distribution (4.1) for illustration purposes.

emphasise that while the response of a neuron to an input is stochastic, as is the case physiologically, we assume that the teacher does not make mistakes.

The goal of learning is to adjust the weights of the network $\boldsymbol{w}$ such that the label predicted by the neuron equals the true label for any input $\boldsymbol{\xi}$, $\sigma = \sigma_\mathrm{T}$. The adaptation of weights is thought to be a main mechanism of memory formation in biological networks [3]. To this end, the neuron needs to infer the teacher $\boldsymbol{T}$. However, the neuron only has indirect access to the teacher via a number of samples $(\boldsymbol{\xi}^\mu, \sigma_\mathrm{T}^\mu)$, where we have now indexed the inputs and their labels with $\mu = 1, \ldots,$ see Fig. 4.1. The exact form of the dynamics will be specified below in Section 4.2. A classic example for neurons performing this kind of associative learning are the Purkinje cells in the cerebellum [94, 95, 110, 111].

4.2 Dynamics

Let us now describe the dynamics of the weights learning a rule from a fixed teacher $\boldsymbol{T}$. Initially, all the weights are independent of each other and in equilibrium in

Algorithm	$\mathcal{F}(\lvert\boldsymbol{w}\rvert, \boldsymbol{w}\cdot\boldsymbol{\xi}^\mu, \sigma_{\mathrm{T}}^\mu)$	*Ref.*
■ Hebbian	1	[97, 106]
■ Perceptron	$\theta(-\sigma_{\mathrm{T}}^\mu \boldsymbol{w}\cdot\boldsymbol{\xi}^\mu)$	[107, 112]
■ AdaTron	$\boldsymbol{w}\cdot\boldsymbol{\xi}^\mu/\sqrt{N}\,\theta(-\sigma_{\mathrm{T}}^\mu \boldsymbol{w}\cdot\boldsymbol{\xi}^\mu)$	[108]

Table 4.1 | Different learning algorithms for a neuron with weights $\boldsymbol{w}$ online-learning a sample $(\sigma_{\mathrm{T}}^\mu, \boldsymbol{\xi}^\mu)$ together with the colour code used throughout the thesis. Here, $\theta(\cdot)$ is the Heaviside step function. References are given to where the algorithm first appeared in a discussion of (the statistical mechanics of) neural learning, to the best of our knowledge. A detailed discussion of the form of these algorithms is given in Sec. 4.4

the potential

$$V(\boldsymbol{w}) = \frac{k}{2}\boldsymbol{w}\cdot\boldsymbol{w} \tag{4.5}$$

which restricts the weights from increasing indefinitely, as is also the case physiologically [96].

Starting at time $t = 0$, the weights w_n obey overdamped Langevin equations [73]

$$\dot{w}_n(t) = F_n\left(\boldsymbol{w}(t), \sigma_{\mathrm{T}}^{\mu(t)}, \boldsymbol{\xi}^{\mu(t)}, t\right) + \zeta_n(t) \tag{4.6}$$

as in our previous chapter, *cf.* Eq. (3.31). The thermal white noise $\zeta_n(t)$ has correlations $\langle\zeta_n(t)\zeta_m(t')\rangle = 2D\delta_{nm}\delta(t-t')$, where D is the "diffusion" constant. We set the mobility of the weights to unity and impose the fluctuation-dissipation relation $\beta D = 1$ for thermodynamic consistency [20]. We still use angled brackets $\langle\cdot\rangle$ to indicate averages over the thermal noise, unless indicated otherwise.

The total force $\boldsymbol{F} = (F_1, \ldots, F_N)$ on the weights has a conservative contribution from the harmonic potential, $-\nabla V(\boldsymbol{w}) = -k\boldsymbol{w}$, and a non-conservative contribution from the learning force $\boldsymbol{f}$, which is a function of a single sample $(\sigma_{\mathrm{T}}^{\mu(t)}, \boldsymbol{\xi}^{\mu(t)})$. The learning force changes the weights in such a way that the neuron becomes more likely to predict the true label for the input as discussed above. In this chapter, we focus our discussion on *online learning*, where the learning force changes the weights using just a single sample at a time*. The succession of samples is described by the function $\mu(t) = 1, 2, \ldots$. This function may be deterministic or stochastic and we do not make any assumptions about the rate of change of the inputs nor whether the same input may be shown more than once to the neuron.

* Our results also hold for *batch learning*, where the neuron has simultaneous access to a set of samples at any point in time as discussed in detail in Appendix 4.D.

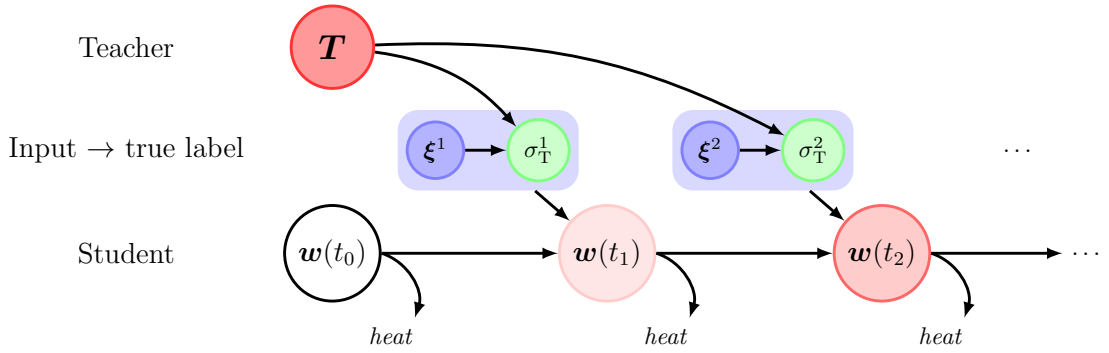


Figure 4.1 | A single neuron learning a rule. The neuron, characterised by its weights $\boldsymbol{w} \in \mathbb{R}^N$, is presented with a succession of inputs $\boldsymbol{\xi}^\mu \in \{-1,1\}^N$ and their true labels $\sigma_{\mathrm{T}}^\mu = \mathrm{sgn}(\boldsymbol{T}\cdot\boldsymbol{\xi}^\mu) = \pm 1$ which are determined by a random, static teacher $\boldsymbol{T} \in \mathbb{R}^N$. The goal of learning is to infer the teacher $\boldsymbol{T}$ by using only the information provided by the samples $(\boldsymbol{\xi}^\mu, \sigma_{\mathrm{T}}^\mu)$, such that the neuron is eventually able to predict the true label of a previously unseen input.

We will assume that the change to the weights in response to a sample is made in the direction of that input, as is the case for most customary algorithms (see Sec. 4.4 and [8, 92, 113]). We thus write $\boldsymbol{f} = (f_1, \ldots, f_N)$ with

$$f_n \equiv \nu(t)\xi_n^{\mu(t)}\sigma_{\mathrm{T}}^{\mu(t)}\mathcal{F}(|\boldsymbol{w}(t)|, \boldsymbol{w}(t)\cdot\boldsymbol{\xi}^{\mu(t)}, \boldsymbol{T}\cdot\boldsymbol{\xi}^{\mu(t)}), \tag{4.7}$$

where we have introduced a possibly time-dependent learning rate $\nu(t)$ * and we denote the Euclidean norm of a vector by $|\cdot|$. Here, $\mathcal{F}$ is an as yet unspecified scalar function of the length of the weight vector, $|\boldsymbol{w}(t)|$, the student's field $\boldsymbol{w}(t)\cdot\boldsymbol{\xi}^{\mu(t)}$ and the teacher's field $\boldsymbol{T}\cdot\boldsymbol{\xi}^{\mu(t)}$. The learning force may only depend on the sign of the teacher's field. Its precise form is specified by learning algorithms; some popular forms are summarised in Tab. 4.1 and described in more detail in Sec. 4.4. However, we stress that the bounds that we derive in this chapter do not depend on the particular form of the learning force and hold for all learning dynamics of the form (4.6). The full Langevin equation for a weight then reads

$$\dot{w}_n(t) = -kw_n(t) + \nu(t)\xi_n^{\mu(t)}\sigma_{\mathrm{T}}^{\mu(t)}\mathcal{F}(|\boldsymbol{w}(t)|, \boldsymbol{w}(t)\cdot\boldsymbol{\xi}^{\mu(t)}, \boldsymbol{T}\cdot\boldsymbol{\xi}^{\mu(t)}) + \zeta_n(t). \tag{4.8}$$

On the ensemble level, the system is fully described by the distribution $p(\boldsymbol{T}, \boldsymbol{w}, t)$. Its equation of motion is given by a Fokker-Planck equation [73] whose form is simplified by the fact that the noise $\zeta_n(t)$ of the different weights is uncorrelated. The

* We repeat that the learning rate that we denote $\nu(t)$ in this chapter is an established concept in the analysis of neural networks and should not be confused with the thermodynamic learning rate l_n, see also Eq. (2.41) and comments thereafter.

dynamics are hence multipartite [38, 84] and the Fokker-Planck equation corresponding to the Langevin dynamics (4.8) separates into one probability current for every weight w_n,

$$\partial_t p(\boldsymbol{T}, \boldsymbol{w}, t) = -\sum_n^N \partial_n j_n(\boldsymbol{T}, \boldsymbol{w}, t), \tag{4.9}$$

where $\partial_t \equiv \partial/\partial t$, $\partial_n \equiv \partial/\partial w_n$ and the probability currents are given by

$$j_n(\boldsymbol{T}, \boldsymbol{w}, t) = \left[-kw_n + \nu(t)\xi_n^{\mu(t)}\sigma_{\mathrm{T}}^{\mu(t)}\mathcal{F}(|\boldsymbol{w}|, \boldsymbol{w}\cdot\boldsymbol{\xi}^{\mu(t)}, \boldsymbol{T}\cdot\boldsymbol{\xi}^{\mu(t)})\right] p(\boldsymbol{T}, \boldsymbol{w}, t) - D\partial_n p(\boldsymbol{T}, \boldsymbol{w}, t). \tag{4.10}$$

There are hence three sources of stochasticity in the system. On the one hand, the fluctuating weights $w(t)$ and the stochastic process of firing an action potential or not, σ, for a given activation (4.2) affect the performance of the network. Furthermore, there is randomness in the choice of samples during learning. Since the neuron learns using just a single randomly drawn input and its label at a time, the system performs stochastic gradient descent in the sense that the direction of the learning force fluctuates from one input to the next and only yields the appropriate direction for the weights on average.

4.3 A first thermodynamic bound on generalising

The aim of the neuron is to predict the label of a previously unseen input $\boldsymbol{\xi}$ as well as possible. In the following discussion, we consider the generalisation properties of the neuron, *i.e.* its performance on an input drawn at random from the distribution (4.1), so we drop the superscript μ on inputs and labels. We quantify the accuracy of the predictions using information theory [8, 31]. The natural quantity to measure the information learnt by the neuron is the mutual information

$$I(\sigma_{\mathrm{T}} : \sigma) \equiv S(\sigma_{\mathrm{T}}) - S(\sigma_{\mathrm{T}}|\sigma) \geq 0 \tag{4.11}$$

which measures by how much, on average, the uncertainty about σ_{T} is reduced by knowing σ for any input. If learning and predicting went perfectly, then by knowing the neuron's output σ one could predict the true label σ_{T} with perfect accuracy, such that $S(\sigma_{\mathrm{T}}|\sigma) = 0$ and hence $I(\sigma_{\mathrm{T}} : \sigma) = \ln 2$. On the other hand, when the weights are in equilibrium in their potential $V(\boldsymbol{w})$ before learning, there is no correlation between the weights of the student and those of its teacher, such that $I(\sigma_{\mathrm{T}} : \sigma) = 0$.

We can connect the mutual information $I(\sigma_{\mathrm{T}} : \sigma)$ to the well-established generalisation error ϵ of neural networks [92, 113]. It gives the probability that the neuron predicts the wrong label for an arbitrary input $\boldsymbol{\xi}$, assuming that the prediction of the neuron is noise-free, *i.e.* $\sigma = \mathrm{sgn}(\boldsymbol{w} \cdot \boldsymbol{\xi})$, and is defined as

$$\epsilon = \langle \theta(-\boldsymbol{w} \cdot \boldsymbol{T}) \rangle \tag{4.12}$$

where θ is the Heaviside step function. If the neuron predicted a label based on its activity reliably via $\sigma = \mathrm{sgn}(\boldsymbol{w} \cdot \boldsymbol{\xi})$ like the teacher, Eq. (4.4), the mutual information between the true and predicted label for an arbitrarily drawn input could be expressed as

$$I(\sigma_{\mathrm{T}} : \sigma) = \ln 2 - S(\epsilon), \tag{4.13}$$

where $S(p) = -p \ln p - (1-p) \ln(1-p)$ is the shorthand for the Shannon entropy of a binary stochastic variable with probability p. For a realistic neuron, the activity gives only the probability that the neuron will fire an action potential, see Eq. (4.3), hence Eq. (4.13) constitutes an upper bound on its actual performance with noisy predictions. In the following, we will focus on deriving thermodynamic bounds on the amount of information that the neuron can learn from its teacher for the ideal case of noise-free predictions.

Thermodynamics enters the picture by considering the free energy costs of the non-equilibrium dynamics of the weights. Similarly to Chapter 3, they can be quantified by the total entropy production $\Delta S_n^{\mathrm{tot}}$ of a single weight in the network which is guaranteed to be non-negative by the second law of stochastic thermodynamics and has two contributions: the heat dissipated by the nth weight into the connected heat bath, ΔQ_n, and the change in Shannon entropy of the marginalised distribution $p(w_n)$ [20].

For a neural network learning with the dynamics (4.8), we can show both for $N = 1$ and in the thermodynamic limit that

$$I(\sigma_{\mathrm{T}} : \sigma) \leq \Delta S_n^{\mathrm{tot}} \equiv \Delta S(w_n) + \Delta Q_n \tag{4.14}$$

from the second law for the network (see Appendices 4.A and 4.B for details). This suggests the introduction of an efficiency

$$\eta \equiv \frac{I(\sigma_{\mathrm{T}} : \sigma)}{\Delta S(w_n) + \Delta Q_n} \leq 1. \tag{4.15}$$

This inequality is our first main result and holds at all times $t > 0$ in Eq. (4.8) and (4.9).

We note that while this result is superficially similar to the inequality we have derived in Chapter 3, here we consider an entirely different scenario. In the previous chapter, there was no teacher; instead, we considered the learning of a number

of *fixed* inputs with true labels drawn at random, such that the true labels were uncorrelated to the inputs and to each other. Hence the concept of a generalisation error did not apply and "the information" was always related to the labels of the fixed set of inputs. Here, we are learning from a number of samples $(\sigma_{\mathrm{T}}^{\mu}, \boldsymbol{\xi}^{\mu})$ which are examples of the function that the teacher performs, Eq. (4.4). The network tries to infer this function in order to be able to correctly classify *previously unseen* inputs. What we show here is that the ability to learn from a teacher and generalise accordingly is bound by the total entropy production per weight. We will come back to the differences between the learning problems considered in this and the previous chapter in Sec. 4.7.

4.4 Efficiency of different learning algorithms with $N = 1$

Let us look at a toy model of a neuron with a single weight. The weight is initially in equilibrium in the harmonic potential $V(w) = kw^2/2$. Without loss of generality, we can set $k = \beta = D = 1$, making energy, entropy, time and the weights dimensionless. At time $t = 0$, the learning rate is suddenly increased from 0 to a constant value ν_0.

The neuron learns using one of the three learning algorithms, each defined by a particular choice of $\mathcal{F}$ and summarised in Tab. 4.1. The simplest non-trivial choice is $\mathcal{F} = 1$, which is Hebbian learning [97, 106]. For such an algorithm, each incoming sample changes the weight by an amount $\sim \sigma_{\mathrm{T}}^{\mu}\xi^{\mu}$. An obvious improvement on this simple algorithm is to only change the weight if the network would currently predict the wrong label for that input, which is achieved by choosing $\mathcal{F} = \theta(-\sigma_{\mathrm{T}}^{\mu} w\xi^{\mu})$. This is the Perceptron algorithm [107]. A further refinement of this rule is achieved by choosing $\mathcal{F} = |w\xi^{\mu}|\theta(-\sigma_{\mathrm{T}}^{\mu} w\xi^{\mu})$ such that the change in the weights is proportional to the confidence of the neuron in its decision, measured by $|w\xi^{\mu}|$.

The key insight to solve the dynamics in each case is that in one dimension, $\sigma_{\mathrm{T}}^{\mu}\xi^{\mu} = \operatorname{sgn}(T)$ for all ξ^{μ}, which is readily verified. This has the appealing consequence that it is possible to rewrite the Langevin equations for all three learning rules without any mention of the inputs ξ^{μ}. Instead, learning a rule is equivalent to a quench of the potential of the weight from the simple harmonic form $V(w) = w^2/2$ to a new T-dependent potential $V^{\mathrm{q}}(T, w)$, the exact form of which depends on the learning algorithm chosen. They read

$$V^{\mathrm{q}}(T, w) = \begin{cases} w^2/2 - \nu_0 w \operatorname{sgn}(T) \\ w^2/2 - \nu_0 w \operatorname{sgn}(T)\theta(-wT) \\ w^2/2(1 - \nu_0 \operatorname{sgn}(T)\operatorname{sgn}(w)\theta(-w \operatorname{sgn}(T))) \end{cases} \tag{4.16}$$

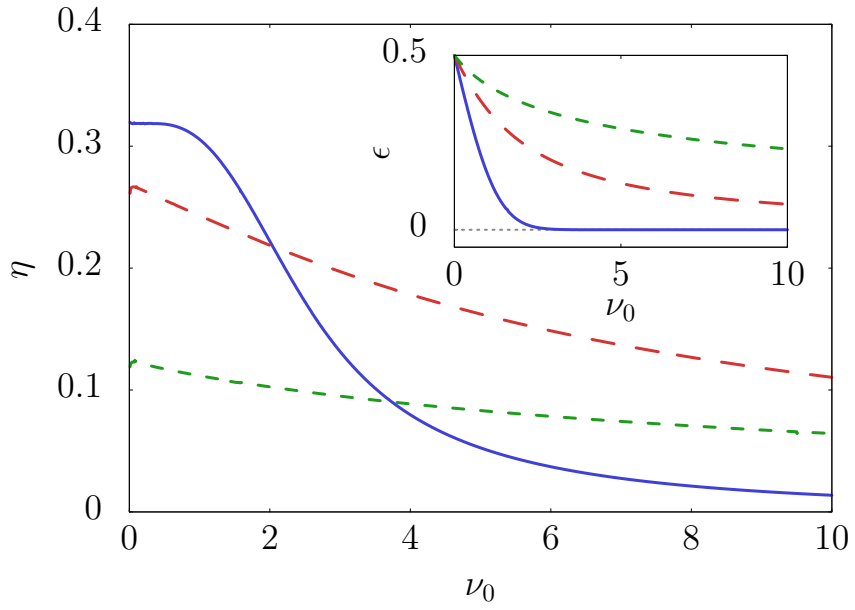


Figure 4.1 | Efficiency of a toy model with $N = 1$. We plot the efficiency η, Eq. (4.15) and in the inset the generalisation error ϵ, Eq. (4.12), as a function of the fixed learning rate ν_0 for a neuron with a single weight, $N = 1$, learning using the Hebbian ■ (solid), Perceptron ■ (long dashed) and AdaTron ■ (dashed) algorithms. Parameters: $k = \beta = D = 1$ without loss of generality.

for Hebbian, Perceptron and AdaTron learning, respectively. The weight then relaxes to the new equilibrium distribution, which is given by the Boltzmann distribution. The heat dissipated by the weight during this isothermal relaxation is given by

$$\Delta Q = \langle V^{\mathrm{q}}(T, w)\rangle_0 - \langle V^{\mathrm{q}}(T, w)\rangle_{\mathrm{eq}} \tag{4.17}$$

where $\langle\cdot\rangle_0$ and $\langle\cdot\rangle_{\mathrm{eq}}$ indicate averages with respect to the distributions of teacher and weight at $t = 0$ and after relaxation, respectively.

We plot the efficiency of learning (4.14) for $t \to \infty$ in Fig. 4.1 as a function of the learning rate. While the Hebbian algorithm yields the lowest generalisation error, its efficiency is quickly dominated by the heat dissipated, $\Delta Q \sim \nu_0^2$, resulting in low efficiency. The perceptron algorithm is the most efficient for large ν and yields a better generalisation performance than the AdaTron algorithm, too (see the inset of Fig. 4.1).

We finally note that our inequality (4.14) is sharp for $N = 1$. Optimal efficiency $\eta \to 1$ can for example be reached for Hebbian learning with a time-dependent learning rate $\nu(t)$ where we first linearly increase $\nu(t)$ from 0 to ν_0 over a period

of time τ and then keep it at the final value ν_0:

$$\nu(t) \equiv \begin{cases} \nu_0 t/\tau & t < \tau \\ \nu_0 & t \geq \tau, \end{cases} \tag{4.18}$$

which is similar to an example discussed in [91]. In the limit of slow driving $\tau \to \infty$, the dissipated heat $\Delta Q \to 0$. If additionally the learning rate $\nu \to \infty$, the efficiency $\eta \to 1$.

4.5 Learning in large networks and a second bound

We just saw in Sec. 4.4 that for $N = 1$, $\xi^\mu \operatorname{sgn}(T\xi^\mu) = \operatorname{sgn}(T)$ which simplifies the analysis because the inputs ξ^μ do not appear explicitly in the equation of motion of the weight. In higher dimensions, we have instead a learning force on the nth weight

$$f_n \sim \xi_n^\mu \operatorname{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi}^\mu) = \xi_n^\mu \operatorname{sgn}\left(T_n \xi_n^\mu + \sum_{m \neq n} T_m \xi_m^\mu \right) \tag{4.19}$$

which will fluctuate between the desired value $\operatorname{sgn}(T_n)$ and $-\operatorname{sgn}(T_n)$ due to the second term inside the sign function, which is effectively a noise term corrupting the signal from the n-th component of the teacher. So instead of relaxing to a new equilibrium as seen for $N = 1$, the weights relax to a steady state with a constant, positive rate of thermodynamic entropy production [20]. Our inequality (4.14) still applies to this process, but it is not very sharp anymore: $I(\sigma_\mathrm{T} : \sigma) \sim 1$ and $\Delta S(w_n) \sim 1$, but a steady state comes with a non-zero rate of heat dissipation, such that $\Delta Q \sim t$. This issue was not addressed in our previous work [91]. In this section, we derive a sharper bound using concepts from steady state thermodynamics [86].

We start with the explicit expression for the total entropy production of the weights of the network [20],

$$\dot{S}^\mathrm{tot}(t) = \sum_n \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\, \frac{p(\boldsymbol{T}, \boldsymbol{w}, t)}{D} \left(\frac{j_n(\boldsymbol{T}, \boldsymbol{w}, t)}{p(\boldsymbol{T}, \boldsymbol{w}, t)} \right)^2 \geq 0. \tag{4.20}$$

In our problem, the learning rate $\nu(t)$ acts as a control parameter. For every value of $\nu(t)$, there is a well-defined steady state $p^\mathrm{s}(\boldsymbol{T}, \boldsymbol{w}; \nu(t))$ where $\partial_t p^\mathrm{s}(\boldsymbol{T}, \boldsymbol{w}; \nu(t)) = 0$ as in equilibrium, but where at least some of currents $j_n^\mathrm{s}(\boldsymbol{T}, \boldsymbol{w}; \nu(t)) \neq 0$, leading, *inter alia*, to a constant rate of total entropy production $\dot{S}^\mathrm{tot} \geq 0$ in the steady state. By writing $\partial_t p^\mathrm{s}(\boldsymbol{T}, \boldsymbol{w}; \nu(t)) = 0$, we are referring to the hypothetical scenario where the system at time t is compared to a system that is allowed to relax to the

steady state corresponding to the value of the learning rate ν at that time. For the remainder of this section, we will use the shorthands

$$p = p(\boldsymbol{T}, \boldsymbol{w}, t), \qquad p^{\mathrm{s}} = p^{\mathrm{s}}\left(\boldsymbol{T}, \boldsymbol{w}; \nu(t)\right), \tag{4.21}$$

$$j_n = j_n(\boldsymbol{T}, \boldsymbol{w}, t), \qquad j_n^{\mathrm{s}} = j_n^{\mathrm{s}}\left(\boldsymbol{T}, \boldsymbol{w}; \nu(t)\right), \tag{4.22}$$

to keep our notation slim. We can rewrite the total entropy production using concepts from steady state thermodynamics [86, 87, 90] as we discussed in Section 2.5

$$\dot{S}^{\mathrm{tot}}(t) = \dot{S}^{\mathrm{na}}(t) + \dot{S}^{\mathrm{a}}(t) \tag{4.23}$$

where we have introduced the non-adiabatic entropy production

$$\dot{S}^{\mathrm{na}}(t) \equiv \sum_n \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ \frac{p}{D} \left(\frac{j_n}{p} - \frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \geq 0 \tag{4.24}$$

and the adiabatic entropy production

$$\dot{S}^{\mathrm{a}}(t) \equiv \sum_n \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ \frac{p}{D} \left(\frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \geq 0 \tag{4.25}$$

Both entropy production rates are evidently positive. They each correspond to a possible mechanism that leads to the breaking of time symmetry and hence to dissipation: the application of non-equilibrium constraints ($\dot{S}^{\mathrm{a}}$) and the presence of driving ($\dot{S}^{\mathrm{na}}$).

The non-adiabatic entropy production of the system can be written as [90]

$$\dot{S}^{\mathrm{na}}(t) = -\int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ \dot{p}(\boldsymbol{T}, \boldsymbol{w}, t) \ln \frac{p(\boldsymbol{T}, \boldsymbol{w}, t)}{p^{\mathrm{s}}(\boldsymbol{T}, \boldsymbol{w}, \nu(t))}. \tag{4.26}$$

It becomes identically zero once the steady state is reached, as is easily seen from its definition. By splitting the logarithm, we find the second law of steady state thermodynamics [86, 87, 90]

$$\dot{S}^{\mathrm{na}}(t) = \dot{S}(\boldsymbol{T}, \boldsymbol{w}, t) + \dot{Q}^{\mathrm{ex}}(t) \geq 0 \tag{4.27}$$

where $\dot{S}(\boldsymbol{T}, \boldsymbol{w}, t)$ is the rate of change of the Shannon entropy of the distribution $p(\boldsymbol{T}, \boldsymbol{w}, t)$ and we have identified the excess heat [86, 87, 90]

$$\dot{Q}^{\mathrm{ex}}(t) \equiv \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ \dot{p}(\boldsymbol{T}, \boldsymbol{w}, t)\ \ln p^{\mathrm{s}}(\boldsymbol{T}, \boldsymbol{w}, \nu(t)) \tag{4.28}$$

which has no definite sign.

Starting from the second law of steady-state thermodynamics (4.27), we can derive our second, sharper bound on the accuracy of learning:

$$I(\sigma_{\mathrm{T}} : \sigma) \leq \Delta S(w_n) + \Delta Q_n^{\mathrm{ex}} \tag{4.29}$$

which leads to the efficiency

$$\tilde{\eta} \equiv \frac{I(\sigma_{\mathrm{T}} : \sigma)}{\Delta S(w_n) + \Delta Q_n^{\mathrm{ex}}} \leq 1. \tag{4.30}$$

This the second main result of this chapter. It also holds at all times and applies to any learning algorithm that depends on the weights, $\boldsymbol{w}$, and samples $(\sigma_{\mathrm{T}}^{\mu(t)}, \boldsymbol{\xi}^{\mu(t)})$. We give the details of its derivation in Appendix 4.C and show that our result applies to batch learning in Appendix 4.D.

4.6 Online learning in large networks

The number of afferent connections to a single neuron in a realistic network may be on the order of thousands [3], so it is sensible to analyse learning in the limit $N \to \infty$. We will focus on online learning [112, 114] using the algorithms introduced in Sec. 4.4 and summarised in Table 4.1. We will assume that the samples, indexed by $\mu(t)$, change much faster than the weights relax. This assumption is central to virtually all of the existing literature on the analysis of online learning algorithms.

4.6.1 Scaling of the learning rate

We have noted that for $N > 2$, the learning force on the nth weight will fluctuate between two values proportional to $\pm\,\mathrm{sgn}(T_n)$, leading to a steady state with constant ϵ and constant rate of heat dissipation. Let us try to make this statement more quantitative by looking at the learning force averaged over the inputs $\boldsymbol{\xi}$ in the limit $N \to \infty$. Setting $\mathcal{F} = 1$ for the moment for simplicity of notation, we have

$$f_n = \nu(t)\sigma_{\mathrm{T}}^{\mu}\xi_n^{\mu} = \nu(t)\xi_n^{\mu}\,\mathrm{sgn}(T_n\xi_n^{\mu} + \psi) \tag{4.31}$$

where we have written $\mu = \mu(t)$ to simplify our notation and we have introduced the noise term inside the $\mathrm{sgn}(\cdot)$ function,

$$\psi \equiv \sum_{m \neq n} T_m \xi_m^{\mu}. \tag{4.32}$$

ψ is uncorrelated with $T_n\xi_n^{\mu}$ and normally distributed with zero mean and variance $N - 1 \approx N$ due to the central limit theorem since the teacher and the inputs are uncorrelated. We are interested in the probability $p_{\parallel}$ that

$$\mathrm{sgn}(T_n\xi_n^{\mu} + \psi) = \mathrm{sgn}(T_n\xi_n^{\mu}), \tag{4.33}$$

i.e. the probability that the learning force points in the right direction despite the noise term ψ. This probability is found by integrating the binormal distribution $p(T_n, \psi) = p(T_n)p(\psi)$ over the region where (4.33) holds for $\xi_n^\mu = 1$ and $\xi_n^\mu = -1$, respectively. We find that

$$\langle f_n \rangle_{\boldsymbol{\xi}} = \nu(t)(2p_\parallel - 1)\,\mathrm{sgn}(T_n) \sim \nu(t)\frac{\mathrm{sgn}(T_n)}{\sqrt{N}} \tag{4.34}$$

where we have expanded $p_\parallel$ for large N [115]. Hence the larger the network, the smaller the information that $\sigma_\mathrm{T} = \mathrm{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi})$ carries about a single component of the teacher network. This analysis suggests we choose a learning rate $\nu(t) \equiv \tilde{\nu}_0\sqrt{N}$ with the normalised learning rate $\tilde{\nu}(t) \sim 1$. This choice corresponds to the conventional scaling of time with the inverse of the network size in the machine learning literature [92, 113], which amounts to nothing more but an increase in samples shown to the network to compensate for the dilution of the signal.

4.6.2 Dynamics

First of all, we would like to compute the time-dependent generalisation error $\epsilon(t)$ for online learning with the three algorithms from Tab. 4.1 in a large network with dynamics given by (4.8). We keep the inverse temperature and the diffusion constant at $\beta = D = 1$ and again consider the case where the learning rate is quenched to a constant value $\tilde{\nu} = \tilde{\nu}_0$ at $t = 0$, leaving us with two free parameters: $\tilde{\nu}_0$ and the stiffness of the harmonic potential k, see Eq. (4.5).

We thus introduce two new parameters, which go back to the original proof of convergence of the perceptron algorithm [116] and play an important role in the statistical mechanics of learning [92],

$$\mathcal{Q} \equiv \frac{\boldsymbol{w} \cdot \boldsymbol{w}}{N} \quad \text{and} \quad \mathcal{R} \equiv \frac{\boldsymbol{T} \cdot \boldsymbol{w}}{N}. \tag{4.35}$$

These quantities have the appealing property of being self-averaging in the thermodynamic limit, where they become the second moment of w_n and the covariance of (T_n, w_n), respectively. Using geometrical [92] or analytical [113] arguments, it can be shown that the generalisation error (4.12) becomes

$$\epsilon = \frac{1}{\pi}\arccos\left(\frac{\boldsymbol{w} \cdot \boldsymbol{T}}{|\boldsymbol{w}||\boldsymbol{T}|}\right) = \frac{1}{\pi}\arccos\left(\frac{\mathcal{R}}{\sqrt{\mathcal{Q}}}\right). \tag{4.36}$$

Hence it is sufficient to find and solve the equations of motion for $\mathcal{Q}$ and $\mathcal{R}$ to solve the dynamics of the generalisation error.

We can indeed derive such equations directly from the Langevin equation for the weights $\boldsymbol{w}$ (4.8) (see Appendix 4.E). They read

$$\begin{aligned}\dot{\mathcal{Q}} &= 2(1-k\mathcal{Q}) + 2\tilde{\nu}_0 \left\langle \operatorname{sgn}(x) y \mathcal{F}(x,y) \right\rangle_{\boldsymbol{\xi}} \\ &\quad + \tilde{\nu}_0^2 \left\langle \mathcal{F}^2(x,y) \right\rangle_{\boldsymbol{\xi}}, \end{aligned} \tag{4.37a}$$

$$\dot{\mathcal{R}} = -k\mathcal{R} + \tilde{\nu}_0 \left\langle \operatorname{sgn}(x) x \mathcal{F}(x,y) \right\rangle_{\boldsymbol{\xi}}, \tag{4.37b}$$

where we have introduced the auxiliary random variables

$$x \equiv \boldsymbol{T} \cdot \boldsymbol{\xi}/\sqrt{N} \qquad \text{and} \qquad y \equiv \boldsymbol{w} \cdot \boldsymbol{\xi}/\sqrt{N}. \tag{4.38}$$

Since we are assuming that the inputs change on a timescale much faster than the relaxation time of the weights, we need to average Eqs. (4.37) over the inputs $\boldsymbol{\xi}$. This average is simplified by noting that the inputs only enter the equations via x and y. Thus the average over the inputs can be replaced with an average over x and y, which are binormally distributed by the central limit theorem, with moments

$$\begin{aligned} \langle x \rangle_{\boldsymbol{\xi}} = \langle y \rangle_{\boldsymbol{\xi}} = 0, \\ \left\langle x^2 \right\rangle_{\boldsymbol{\xi}} = 1, \quad \left\langle y^2 \right\rangle_{\boldsymbol{\xi}} = \mathcal{Q}, \quad \langle xy \rangle_{\boldsymbol{\xi}} = \mathcal{R}. \end{aligned} \tag{4.39}$$

The averages $\langle \cdot \rangle_{\boldsymbol{\xi}}$ can be performed analytically for all three learning algorithms and their particular choice of $\mathcal{F}$, see Tab. 4.1. We give the results in Appendix 4.E. This procedure eventually yields a set of closed equations for $\mathcal{R}$ and $\mathcal{Q}$ for each learning algorithm, which can be solved numerically.

Fig. 4.1 shows the generalisation error as a function of (scaled) time as obtained from numerical simulations of the Langevin equation (4.8) for a network with $N = 10000$, $\tilde{\nu}_0 = \beta = D = 1$ in dashed lines. The result obtained by our analytical calculation that we just discussed is shown in the same plot using solid lines. First, we note that ϵ is a self-averaging quantity, *i.e.* each simulation run generates the same ϵ over time within small fluctuations which are scale inversely with N. Furthermore, the dynamics of ϵ are well described by our analytical result. While the Hebbian learning takes the longest time to converge, it is perhaps surprisingly the most robust algorithm in the presence of noise, consistently yielding the lowest generalisation errors. Indeed, for online learning with $k = 0$ and no noise, it is well established that ϵ decays slower for the Perceptron than the Hebbian algorithm; on the other hand, Hebbian learning fails miserably with non-uniform input distributions [92]. The performance of the Perceptron is significantly improved by a choice of time-dependent learning rates in a process called *annealing*. This is beyond the scope of this chapter, but see [117] for a detailed discussion of the impact of time-dependent learning rates on the convergence of learning algorithms.

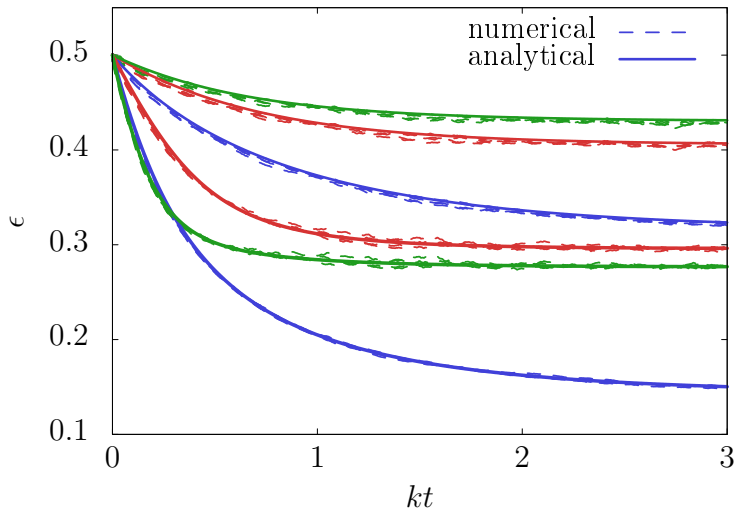


Figure 4.1 | Dynamics of online learning in large networks. We computed ϵ analytically by solving the equations of motion for $\mathcal{Q}$ and $\mathcal{R}$, Eq. (4.37) (solid), and numerically from simulations of the dynamics (4.8) with $N = 10000$ (dashed) with the normalised learning rate $\tilde{\nu}_0 = 1$ in both cases. The bottom three trajectories have $k = 0.1$, while for the top three trajectories, $k = 1$. We compare online learning using the Hebbian ■, Perceptron ■ and AdaTron ■ algorithms, plotting five trajectories for each algorithm. Parameters: $\beta = D = 1$.

Catastrophic failure of AdaTron learning

A remarkable property of AdaTron learning is demonstrated in Fig. 4.2, where we plot the final, steady-state generalisation error ϵ against the normalised learning rate $\tilde{\nu}_0$. While Hebbian and Perceptron learning (green and blue, resp.) show the expected decrease of ϵ with $\tilde{\nu}_0$, there is a sharp increase of ϵ for AdaTron learning at $\tilde{\nu}_c = 3$ (green). Indeed, for large learning rates, the AdaTron algorithm will fail completely. This sensitivity of the algorithm to the value of the learning rate is well-known in the noise-free case without potential $V(\boldsymbol{w})$ [92] and persists in our model with noise, most markedly for low potential stiffness k.

The critical dependence of the generalisation error ϵ on the learning rate $\nu(t) = \sqrt{N}\tilde{\nu}_0$ for AdaTron learning in weak potentials ($k \ll 1$) in the thermodynamic limit $N \to \infty$ is most clearly seen by transforming variables from $(\mathcal{Q}, \mathcal{R})$, Eq. (4.37), to $(\mathcal{Q}, \mathcal{S})$ with

$$\mathcal{S} \equiv \frac{\mathcal{R}}{\sqrt{\mathcal{Q}}}. \tag{4.40}$$

The variables $(\mathcal{Q}, \mathcal{S})$ obey another closed set of equations of motion. After aver-

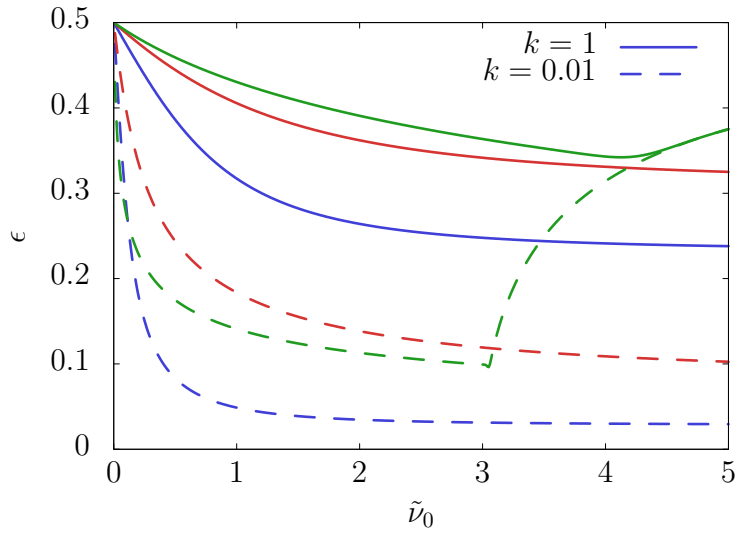


Figure 4.2 | Final generalisation error in large networks. We plot the final, steady-state generalisation error ϵ of online learning in the thermodynamic limit using Hebbian ■, Perceptron ■ and AdaTron ■ algorithms. The behaviour of the algorithms and AdaTron learning in particular is discussed in detail in Sec. 4.6.2. Parameters: $\beta = D = 1$.

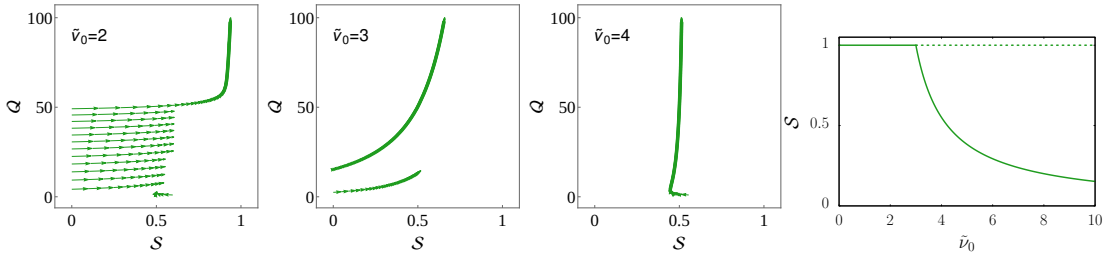


Figure 4.3 | Critical learning rate for AdaTron learning. The first three plots from left to right are vector plots in phase space for the first-order system $(\dot{\mathcal{Q}}, \dot{\mathcal{S}})$, Eq. (4.41), for AdaTron learning in the thermodynamic limit $N \to \infty$ with constant normalised learning rate $\tilde{\nu}(t) = \tilde{\nu}_0$ and $k = 0.01$. For $\tilde{\nu}_0 \leq 2$, there is an attracting state with $\mathcal{S} \to 1$. As we increase the learning rate $\tilde{\nu}_0$, another attracting state appears with $\mathcal{S} \to 1/2$ and hence $\epsilon = 1/2$. In the limit $k \ll 1$, this behaviour can be understood from the bifurcation diagram of the closed, single equation for $\mathcal{S}$, Eq. (4.41b), shown on the far right, where stable (unstable) fixed points are indicated by straight (dashed) lines. Parameters: $\beta = D = 1$.

aging over the inputs using $p(x, y)$ as described in Section 4.E, we find

$$\dot{\mathcal{Q}}(t) = 2 + \mathcal{Q}(t)\left(-2k + (2\tilde{\nu}_0 - \tilde{\nu}_0^2)\frac{\mathcal{S}(t)\sqrt{1-\mathcal{S}(t)^2} - \arccos\mathcal{S}(t)}{\pi}\right), \quad (4.41a)$$

$$\dot{\mathcal{S}}(t) = -\frac{\mathcal{S}(t)}{\mathcal{Q}(t)} + \frac{\tilde{\nu}_0^2\mathcal{S}(t)^2\sqrt{1-\mathcal{S}(t)^2}}{2\pi} + \frac{\tilde{\nu}_0\left(1-\mathcal{S}(t)^2\right)^{3/2}}{\pi} - \frac{\tilde{\nu}_0^2\mathcal{S}(t)\arccos\mathcal{S}(t)}{2\pi}. \quad (4.41b)$$

Three stream plots of this system for $\tilde{\nu}_0 = 2, 3, 4$, shown in Fig. 4.3, reveal a qualitative change in behaviour of the system $(\mathcal{Q}, \mathcal{S})$ away from a solution with $\mathcal{S} \to 1$ and hence $\epsilon \to 0$. Indeed, as $\tilde{\nu}$ increases, $\mathcal{S} \to 0$ and thus $\epsilon \to 1/2$. This observation calls for a more detailed analysis of the system (4.41). Unfortunately, the fixed points of the system cannot be found explicitly. However, we can consider the limit of small k where the transition is most pronounced, see Fig. 4.2. Expanding the equation for $\dot{\mathcal{Q}}(t)$ around $k = 0$ and $\mathcal{S} = 1$ shows that $\mathcal{Q} \sim 1/k$ in the steady state. This suggests neglecting the first term in Eq. (4.41b), which has the appealing consequence of yielding a closed, single equation for $\mathcal{S}(t)$. This equation has a fixed point $\mathcal{S} = 1$, which is easily checked by substitution. Expanding Eq. (4.41b) around $\mathcal{S} = 1$ yields

$$\dot{S}(t) = \frac{2\sqrt{2}}{\pi}\left(\frac{\tilde{\nu}_0^2}{3} - \tilde{\nu}_0\right)(1-\mathcal{S})^{3/2} + \mathcal{O}(1-\mathcal{S})^{5/2} \quad (4.42)$$

from which we see that the derivative will change sign at the critical learning rate $\tilde{\nu}_c = 3$, which is the same value where the well-known breakdown of AdaTron learning occurs for a setup with $k = 0$ and no thermal noise [92]. A detailed graphical analysis reveals that the solution $\mathcal{S} = 1$ looses its stability at $\tilde{\nu}_c = 3$ while a second fixed point emerges, which is stable, leading to the collapse of the generalisation error observed in Fig. 4.2, as shown in the bifurcation diagram for $\mathcal{S}$ in the right-most plot of Fig. 4.3.

4.6.3 Efficiency of learning

We can also derive an ordinary differential equation for the ensemble average of the excess heat (4.28) in terms of $\mathcal{Q}$ and $\mathcal{R}$, with the details to be found in Appendix 4.F. Since the components of the teacher and the weights are normally distributed, the change in Shannon entropy of the marginalised distribution of a weight $\Delta S(w_n)$ can be expressed in terms of just $\mathcal{Q}$, giving us all the information necessary to compute the efficiency of learning $\tilde{\eta}$ (4.30). We plot the efficiency $\tilde{\eta}$ in the thermodynamic limit in Fig. 4.4 against the normalised learning rate $\tilde{\nu}_0$

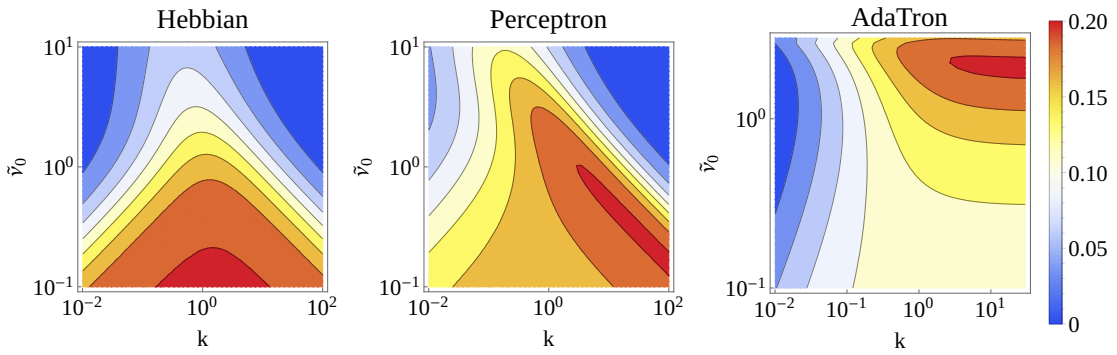


Figure 4.4 | Efficiency of learning in large networks. The efficiency $\tilde{\eta}$, Eq. (4.30), for neural networks performing online learning with fixed normalised learning rate $\tilde{\nu}_0$ using the Hebbian, Perceptron and AdaTron algorithms for online learning in the thermodynamic limit is shown from left to right as a function of the potential stiffness k and $\tilde{\nu}_0$. Parameters: $\beta = D = 1$.

and the potential stiffness k, which are the only remaining free parameters in this model.

The efficiency of Hebbian learning is roughly symmetric with respect to k around $k = 1$, while Perceptron and AdaTron learning display highly asymmetric patterns. However, we find that despite the different patterns, the maximum efficiency for all three algorithms is $\eta \simeq 0.2$. We can dig a little deeper by first noting that since $p(T_n, w_n)$ is normally distributed for the learning algorithms we have considered, both the mutual information $I(T_n : w_n)$ and the mutual information between the true and the predicted label for an arbitrary input $I(\sigma_\mathrm{T} : \sigma)$ can be written as functions of only the correlation between T_n and w_n, $\rho \equiv \mathcal{R}/\sqrt{\mathcal{Q}}$. Expanding around $\rho = 0$ yields

$$\frac{I(\sigma_\mathrm{T} : \sigma)}{I(T_n : w_n)} = \frac{\ln 2 - S(\arccos(\rho)/\pi)}{-1/2 \ln(1-\rho^2)} = \frac{4}{\pi^2} + \mathcal{O}(\rho^2) \simeq 0.4 \tag{4.43}$$

which turns out to be a good approximation for $\rho \lesssim 0.9$. So at maximum efficiency,

$$\frac{I(w_n : T_n)}{\Delta S(w_n) + \Delta Q_n^{\mathrm{ex}}} \simeq \frac{1}{2} \tag{4.44}$$

for all three algorithms.

The plot in Fig. 4.5 shows that the bifurcation we discussed for the AdaTron learning in Sec. 4.6.2 leads to a decaying efficiency $\tilde{\eta} \to 0$ since $I(\sigma_\mathrm{T} : \sigma) \to 0$. This effect is smoothed out with increasing potential stiffness.

Let us finally note that, in this model, the rate of heat dissipation of a single weight diverges in the thermodynamic limit, $\dot{Q}_n \to \infty$ as $N \to \infty$. This is readily

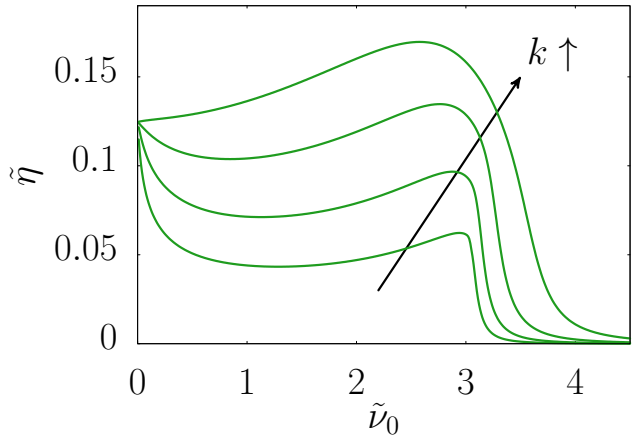


Figure 4.5 | Catastrophic loss of efficiency for AdaTron learning. The efficiency $\tilde{\eta}$, Eq. (4.30), for neural networks performing online learning with fixed normalised learning rate $\tilde{\nu}_0$ using AdaTron learning ■ versus the normalised learning rate $\tilde{\nu}_0$ for increasing values of k from 0.01 (bottom) to 1 (top). We see that the efficiency goes to zero beyond a critical normalised learning rate $\tilde{\nu}_0$. Parameters: $\beta = D = 1$.

understood from a physical point of view since the weights experience a large force, $f \sim \sqrt{N}$, which fluctuates very quickly. This observation reinforces the importance of our second bound involving the excess heat (4.28), which does not diverge even in the limit $N \to \infty$.

4.7 Discussion and perspectives

We have analysed the learning of linearly separable rules by neural networks as a model for the thermodynamics of generalisation. Using stochastic thermodynamics and information theory, we have shown that the accuracy with which the neuron is able to apply the rule to previously unseen inputs is constrained by the dissipation of free energy of a single weight during the learning process. Our results hold for all learning algorithms that have access to all the weights, $\boldsymbol{w}$, and either a set of or a succession of samples $(\sigma_{\mathrm{T}}^{\mu(t)}, \boldsymbol{\xi}^{\mu(t)})$ in batch or online learning, respectively. We have furthermore given a detailed analysis of both the dynamics and the thermodynamics of online learning in large neural networks with noisy dynamics and weights constrained by an external potential.

It is worthwhile to revisit the results of the previous chapter, where we analysed the thermodynamic costs of building a model from data, in the light of these results. In Chapter 3, we studied a different learning problem, namely learning P mappings $\boldsymbol{\xi}^{\mu} \to \sigma_{\mathrm{T}}^{\mu}$ from *fixed* inputs $\boldsymbol{\xi}^{\mu}$, $\mu = 1, \ldots, P$ to their true labels $\sigma_{\mathrm{T}}^{\mu}$. The

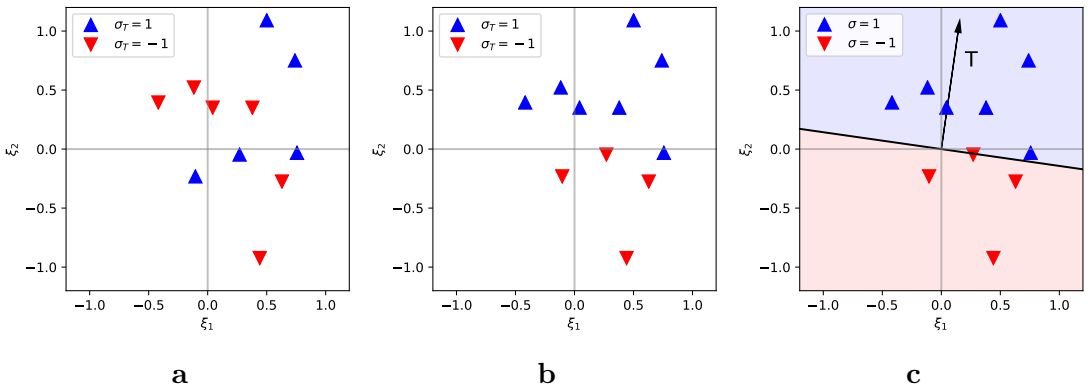


Figure 4.1 | Realisable labellings of random points for a linear perceptron. Random labellings of a fixed set of inputs can sometimes be implemented by a linear perceptron with weights $\boldsymbol{w}$, for which the predicted label σ is only a function of the inner product $\boldsymbol{w} \cdot \boldsymbol{\xi}$. For the fixed set of inputs shown in the three plots above, the random labelling shown in Fig. **a** cannot be realised by a linear perceptron; the labelling in Fig. **b** however can be implemented using the a linear perceptron with the weight vector shown in Fig. **c**

true labels were drawn at random for each input, and hence uncorrelated to the inputs and to each other. Hence there is no generalisation error for this problem – if the true label of every input is determined by pure chance, the mappings $\{\boldsymbol{\xi}^\mu \to \sigma_\mathrm{T}^\mu\}_{\mu=1}^P$ carry no information about the label of a previously unseen input. Instead, the challenge is to find a set of weights that reproduce the mappings faithfully.

The two problems are however related in the following way. It is possible to at least construct a teacher $\boldsymbol{T}$ that reproduces all the mappings $\boldsymbol{\xi}^\mu \to \sigma_\mathrm{T}^\mu$ using $\sigma_\mathrm{T}^\mu = \mathrm{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi}^\mu)$ if and only if the number of mappings P is less than the capacity of the network. This capacity is usually defined in the thermodynamic limit, where the number of weights $N \to \infty$, and we are interested in the relative number of inputs $\alpha_c \equiv P_c/N \sim 1$ for which there exists a teacher $\boldsymbol{T}$ that reproduces all the true labels via $\sigma_\mathrm{T} = \mathrm{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi})$ with probability 1 [8]. Its numerical value can be derived analytically from replica calculations [99], but it was first understood using geometrical arguments [118, 119] (see also [8] for a detailed discussion and Fig. 4.1).

If it is possible to construct a teacher $\boldsymbol{T}$, the rule implicitly defined by the mappings is realisable and can, at least in theory, be learned. Even in that case, however, the issue remains for the scenario considered in [91] that the number of samples from which the neuron learns is limited and might not be sufficient to learn the underlying "rule" effectively. On the other hand, learning the mappings

$\{\boldsymbol{\xi}^\mu \to \sigma_{\mathrm{T}}^\mu\}_\mu^P$ is still a meaningful task even if it is not possible to even construct a network that reproduces them all, if one is willing to accept a certain error in the predictions of the network.

There is still plenty of room for further work along the lines of this chapter. It would be intriguing to consider the generalisation of our model to multi-valued teacher functions, *e.g.* for a network learning to classify digits. The teacher could also be made subject to noise in its outputs σ_{T}, or its components, T_n, or both. Another intriguing learning problem is that of a changing environment, modeled by a drifting teacher [120, 121]. Designing a learning algorithm that optimises the thermodynamic efficiency looks like a serious challenge. More broadly, studying the thermodynamic costs of learning to generalise might form a suitable basis to consider the thermodynamics of decision-making [122].

Appendices to chapter 4

The following appendices give a detailed proof of our main results, inequalities (4.14) and (4.29), in Appendices 4.A to 4.C. Appendix 4.D discusses how our results apply to batch learning. Detailed calculations for the learning dynamics in the thermodynamic limit are given in Appendices 4.E to 4.F.

4.A Derivation of inequality (4.14)

Our first main result, Eq. (4.14), can be derived from the second law of stochastic thermodynamics [20] which states that the rate of total entropy production of the full system is positive

$$\dot{S}^{\mathrm{tot}}(t) \geq 0. \tag{4.45}$$

We will drop the explicit time argument in the following discussion but emphasise that since the distribution $p(\boldsymbol{T}, \boldsymbol{w}, t)$ is time-dependent, so are of course all the quantities derived from it.

Using the results from Section 2.4, we can rewrite the second law for the nth weight as

$$\dot{S}_n^{\mathrm{tot}} = \partial_t S(w_n) + \dot{Q}_n - l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n) \geq 0. \tag{4.46}$$

which is the starting point of our derivation. We note that for the isothermal environment that we assume in this chapter, the rate of thermodynamic entropy production is the heat dissipated into the environment, $\dot{S}_n^{\mathrm{m}} = \dot{Q}_n$, where we remind ourselves that we have set the temperature to unity.

Integrating the N second laws (4.46) with respect to time from $t' = 0$ to $t > 0$ yields

$$\sum_n^N [\Delta S(w_n) + \Delta Q_n] \geq \sum_n^N \int_0^t \mathrm{d}t' \; l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n) \tag{4.47}$$

where we write $\Delta S(w_n)$ and ΔQ_n to denote the total change in Shannon entropy of the distribution $p(w_n)$ and the total heat dissipated by the dynamics of the nth

weight up to time t, respectively. We can interpret the right-hand side of (4.47) by computing the time-derivative of the mutual information $I(\boldsymbol{T} : \boldsymbol{w})$,

$$\partial_t I(\boldsymbol{T} : \boldsymbol{w}) = \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ \left[\partial_t p(\boldsymbol{T}, \boldsymbol{w}, t)\right] \ln \frac{p(\boldsymbol{T}, \boldsymbol{w}, t)}{p(\boldsymbol{T}, t) p(\boldsymbol{w}, t)}. \tag{4.48}$$

Using the Fokker-Planck Eq. (4.9) and integrating by parts, we find

$$\begin{aligned}
\partial_t I(\boldsymbol{T} : \boldsymbol{w}) &= \sum_n^N \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ j_n(\boldsymbol{T}, \boldsymbol{w}, t) \partial_n \ln \frac{p(\boldsymbol{T}, \boldsymbol{w}, t)}{p(\boldsymbol{T}, t) p(\boldsymbol{w}, t)} && (4.49)\\
&= \sum_n^N \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ j_n(\boldsymbol{T}, \boldsymbol{w}, t) \partial_n \ln \frac{p(\boldsymbol{T}, \boldsymbol{w}, t)}{p(w_n, t) p(\overline{\boldsymbol{w}}_n | w_n, t)} && (4.50)\\
&= \sum_n^N l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n) - \sum_n^N \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\ j_n(\boldsymbol{T}, \boldsymbol{w}, t) \partial_n \ln p(\overline{\boldsymbol{w}}_n | w_n, t) && (4.51)\\
&= \sum_n^N l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n) - \left(-\partial_t S(\boldsymbol{w}) + \sum_n^N \partial_t S(w_n) \right) && (4.52)
\end{aligned}$$

where in the penultimate line, we have recovered the integrand on the right-hand side of (4.47). Integrating the term in brackets in Eq. (4.52) with respect to time yields for all times $t > 0$

$$\int_0^t \mathrm{d}t'\ \left(\sum_n^N \partial_{t'} S(w_n) - \partial_{t'} S(\boldsymbol{w}) \right) = \sum_n^N S(w_n) - S(\boldsymbol{w}) \geq 0 \tag{4.53}$$

where we have used that at time $t = 0$, all the weights are independent of each other and hence $S(\boldsymbol{w}) = \sum_n S(w_n)$. The inequality follows from the fact that for any set of random variables, $\sum_n S(w_n) \geq S(\boldsymbol{w})$ [31]. Using this inequality, we can deduce from (4.52) that

$$\sum_n^N \left[\Delta S(w_n) + \Delta Q_n\right] \geq \sum_n^N \int_0^t \mathrm{d}t'\ l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n) \geq I(\boldsymbol{T} : \boldsymbol{w}). \tag{4.54}$$

Using the chain rule of mutual information [31] and the fact that the components

$\boldsymbol{T}$ are independent of each other (Sec. 4.1), we have

$$I(\boldsymbol{T}:\boldsymbol{w}) = \sum_n^N I(T_n:\boldsymbol{w}|T_{n-1},\ldots,T_1) \tag{4.55}$$

$$= \sum_n^N I(T_n:\boldsymbol{w},T_{n-1},\ldots,T_1) - \sum_n^N I(T_n:T_{n-1},\ldots,T_1) \tag{4.56}$$

$$\geq \sum_n^N I(T_n:\boldsymbol{w}) \tag{4.57}$$

$$= \sum_n^N I(T_n:w_n) + I(T_n:\overline{\boldsymbol{w}}_n\,|\,w_n) \tag{4.58}$$

$$\geq NI(T_n:w_n) \tag{4.59}$$

where the inequalities follow again from the non-negativity of mutual information and the fact that all the weights and all the components of the teacher are statistically identical. Using the latter argument for the total entropy production and inserting our last result into the integrated form of the second law (4.47), we find that

$$\Delta S(w_n) + \Delta Q_n \geq I(w_n:T_n) \tag{4.60}$$

Finally, we need to show that the mutual information between the nth component of the weight and teacher vectors are an upper bound on the mutual information between the true and predicted labels of any sample $\boldsymbol{\xi}$,

$$I(w_n:T_n) \geq I(\sigma_{\mathrm{T}}:\sigma). \tag{4.61}$$

Our strategy will be to show that the inequality (4.61) holds even if the neuron predicts a label deterministically, $\sigma = \mathrm{sgn}(\boldsymbol{w}\cdot\boldsymbol{\xi})$. The generalisation error is then the lowest for a given noise level in the weights and given by $I(\sigma_{\mathrm{T}}:\sigma) = \ln 2 - S(\epsilon)$.

Let us first consider a network with $N = 1$. We start by noting that for arbitrary random variables X and Y and an arbitrary function $F(Y)$, we can always write $p(x,y,f(y)) = p(x)p(y|x)p\left(f(y)|y\right)$. We thus identify $X \to Y \to F$ as a Markov chain and find

$$I(X:Y) \geq I(X:F) \tag{4.62}$$

using the data processing inequality [31]. For $N = 1$, we can apply this result twice to show that

$$I(w:T) \geq I(w\xi:T\xi) \geq I(\sigma:\sigma_{\mathrm{T}}) \tag{4.63}$$

as required.

In the thermodynamic limit $N \to \infty$, we use the auxiliary variables $x \equiv \boldsymbol{w} \cdot \boldsymbol{\xi}/\sqrt{N}$ and $y \equiv \boldsymbol{T} \cdot \boldsymbol{\xi}/\sqrt{N}$. We then have from (4.62)

$$I(\sigma_{\mathrm{T}} : \sigma) \leq I(x : y) \tag{4.64}$$

since σ_{T} and σ are functions of x and y, Eq. (4.4) and Eq. (4.3), respectively. We can now average x and y over the inputs (4.1) using $\langle \xi_n \rangle_{\xi} = 0$ and $\langle \xi_n \xi_m \rangle_{\xi} = \delta_{nm}$. By the central limit theorem, x and y are then distributed according to a bivariate Gaussian distribution with correlation [31]

$$\rho \equiv \frac{\mathrm{cov}(w_n, T_n)}{\mathrm{sd}(w_n)\,\mathrm{sd}(T_n)} = \frac{\boldsymbol{w} \cdot \boldsymbol{T}}{|\boldsymbol{w}||\boldsymbol{T}|} \tag{4.65}$$

This is a crucial step in our derivation since it allows us to connect the statistics of teacher and weight in one dimension to the statistics of the true and predicted labels, which are functions of the vectors $\boldsymbol{T}$ and $\boldsymbol{w}$. The mutual information of two variables with a bivariate Gaussian distribution is a function of their correlation alone [31],

$$I_G(w_n : T_n) = -\frac{1}{2}\left(1 - \ln \rho^2\right) = I(x : y) \tag{4.66}$$

which would *also* be the mutual information $I(w_n : T_n)$ if w_n and T_n were jointly distributed normally, which they are not necessarily. However, we can show that $I_G(w_n : T_n)$ is a lower bound on $I(w_n : T_n)$ using the maximum entropy principle. This is a prescription for finding the probability distribution that maximises the Shannon entropy given a number of constraint on the distribution, usually in the form of fixed moments. We briefly review this concept in Appendix 4.B. The crucial point here is that a Gaussian distribution is the maximum entropy distribution for a given covariance matrix. We will denote the maximum entropy notations with an asterisk, *e.g.* p^*.

The mutual information $I(w_n : T_n)$ can be expressed as the relative entropy or Kullback-Leibler divergence between the joint distribution $p(w_n, T_n)$ and the factorised distribution $p(T_n)p(w_n)$ [31]:

$$I(w_n : T_n) = D\left[p(T_n, w_n) \parallel p(T_n)p(w_n)\right] \tag{4.67}$$

$$\equiv \int \mathrm{d}\boldsymbol{T}\,\mathrm{d}\boldsymbol{w}\; p(\boldsymbol{T}, \boldsymbol{w}) \ln \frac{p(\boldsymbol{T}, \boldsymbol{w})}{p(\boldsymbol{T})p(\boldsymbol{w})} \tag{4.68}$$

where the inequality is true for arbitrary probability distributions. Introducing

the shorthand $p(w_n) \equiv p_w$ etc. to simplify notation, we hence are left to show that

$$
\begin{aligned}
&I(w_n : T_n) - I_G(w_n : T_n) \\
&= \left\langle \ln \frac{p_{Tw}}{p_T p_w} \right\rangle_p - \left\langle \ln \frac{p^*_{Tw}}{p^*_T p^*_w} \right\rangle_{p^*} \\
&= \langle \ln p_{Tw} \rangle_p - \langle \ln p^*_{Tw} \rangle_{p*} \\
&\quad - \left[\langle \ln p_T \rangle_p - \langle \ln p^*_T \rangle_{p^*} + \langle \ln p_w \rangle_p - \langle \ln p^*_w \rangle_{p^*} \right] \\
&= \langle \ln p_{Tw} \rangle_p - \langle \ln p^*_{Tw} \rangle_p \\
&\quad - \left[\langle \ln p_T \rangle_p - \langle \ln p^*_T \rangle_p + \langle \ln p_w \rangle_p - \langle \ln p^*_w \rangle_p \right] \\
&= D\left[p_{Tw} \parallel p^*_{Tw} \right] - D\left[p_T \parallel p^*_T \right] - D\left[p_w \parallel p^*_w \right] \\
&= D\left[p_{T|w} \parallel p^*_{T|w} \right] \geq 0
\end{aligned}
$$

where we used that $\langle \ln p^*_{Tw} \rangle_{p^*} = \langle \ln p^*_{Tw} \rangle_p$ for the third equality, see Sec. 4.B, while for the last equality we applied the chain rule for the Kullback-Leibler distance [31] and remembered that $p(T_n)$ is a Gaussian distribution and hence the maximum entropy distribution for a given variance. This completes our derivation of the bound (4.14).

4.B Surprise and maximum entropy distributions

We briefly review the concept of a maximum entropy distributions, which have a long history in physics [123]. We will focus on the case of a single variable to illustrate the concepts, but we note that the multi-dimensional case can be treated using the same methods.

We are looking for a probability distribution p of a continuous variable X with support $\mathcal{S}$ which is subject to M constraints, namely

$$
\int_{\mathcal{S}} \mathrm{d}x \; p(x) r_i(x) = \alpha_i. \tag{4.69}
$$

The maximum entropy prescription for finding the distribution $p(x)$ is to find the distribution that maximises the Shannon entropy $S(X)$ (2.19) under the constraints (4.69). This is a standard calculation using variational calculus; here we

simply quote the result [31],

$$p^*(x) \sim \exp\left(\lambda_0 - 1 + \sum_i^M \lambda_i r_i(x)\right) \tag{4.70}$$

where λ_i are the Lagrange multipliers chosen such that $p(x)$ obeys the constraints. Proving that (4.70) is a maximum is a rather involved calculation and is more easily proven using information-theoretic methods [31].

The key point for our purposes is that for a distribution $p(x)$, which is unknown except for a number of its moments, averaging the surprise $\ln p^*(x)$ of the maximum entropy distribution for the known moments over p^* is equal to the average taken with respect to the true distribution p,

$$\langle \ln p^*(x) \rangle_{p^*} = \langle \ln p^*(x) \rangle_p \,. \tag{4.71}$$

This result is a direct consequence of the form of the maximum entropy distribution and the fact that the moments that enter p^* are by construction equal to the corresponding moments of p.

4.C Derivation of inequality (4.29)

The non-adiabatic entropy production of the nth single weight is defined as

$$\dot{S}_n^{\mathrm{na}}(t) \equiv \int \mathrm{d}\boldsymbol{T}\, \mathrm{d}\boldsymbol{w}\; p \left(\frac{j_n}{p} - \frac{j_n^{\mathrm{s}}}{p^{\mathrm{s}}} \right)^2 \geq 0, \tag{4.72}$$

see Sec. 4.5 for a detailed discussion. Summing $\dot{S}_n^{\mathrm{na}}(t)$ over all subsystems, we find

$$\sum_n \dot{S}_n^{\mathrm{na}}(t) = \sum_n \dot{S}_n(\boldsymbol{T}, \boldsymbol{w}) + \sum_n \dot{Q}_n^{\mathrm{ex}} \geq 0 \tag{4.73}$$

where the rate of excess heat production $\dot{Q}_n^{\mathrm{ex}}$ was defined in Eq. (4.28). After writing $\dot{S}_n(\boldsymbol{T}, \boldsymbol{w}) = \partial_t S(w_n) - l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n)$ and integrating over time, see the discussion in Appendix 4.A, we find that

$$\sum_n^N [\Delta S(w_n) + \Delta Q_n^{\mathrm{ex}}] \geq \sum_n^N \int_0^t \mathrm{d}t'\; l_n(w_n : \boldsymbol{T}, \overline{\boldsymbol{w}}_n). \tag{4.74}$$

and we can now proceed along the lines of Appendix 4.A.

4.D Batch learning

Our discussion has focused on online learning, where, at any one point in time, the network experiences a learning force $\boldsymbol{f}$ due to a single input and its label, Eq. (4.7). Another approach is to average the learning force over a set $D = \{(\sigma_{\mathrm{T}}^{\mu}, \boldsymbol{\xi}^{\mu})\}_{\mu=1}^{P}$ of P inputs and their labels,

$$f_n \equiv \nu(t) \left\langle \xi_n^{\mu} \sigma_{\mathrm{T}}^{\mu} \mathcal{F}(|\boldsymbol{w}(t)|, \boldsymbol{w}(t)\cdot\boldsymbol{\xi}^{\mu}, \boldsymbol{T}\cdot\boldsymbol{\xi}^{\mu}) \right\rangle_D . \tag{4.75}$$

This strategy is called *batch learning* and P is usually chosen to be on the order of N. In the thermodynamic limit, as $N \to \infty$ one thus lets $P \to \infty$ while keeping the ratio $\alpha \equiv P/N$ on the order of one.

Batch learning clearly comes with high requirements in terms of memory. It is generally more efficient than online learning, although the latter can achieve generalisation errors which at least asymptotically match the results from batch learning [92].

Our two main results, inequalities (4.14) and (4.29) apply to batch learning as well. This is because in our derivation, we only used the fact that the teacher enters the force on the weights, albeit indirectly. We did not have to specify the exact form of the learning force that introduces the correlations between the weight and the teacher. Hence it does not make a difference in the derivation of the inequalities whether the learning force is computed for just a single sample or averaged over a set of samples.

4.E Solving the learning dynamics in the thermodynamic limit

Here we give a detailed derivation of the equations of motion for the order parameters $\mathcal{Q}$ and $\mathcal{R}$ introduced in Sec. 4.6 in the thermodynamic limit $N \to \infty$. These are most easily derived by rewriting the Langevin equations for the weights, Eq. (4.8), as Itô stochastic differential equations [74] which we introduced in Sec. 2.1. Here, it reads

$$\mathrm{d}\boldsymbol{w}(t) = -k\boldsymbol{w}(t)\,\mathrm{d}t + \nu(t)\boldsymbol{\xi}^{\mu(t)}\sigma_{\mathrm{T}}^{\mu(t)}\mathcal{F}(|\boldsymbol{w}(t)|, \boldsymbol{w}(t)\cdot\boldsymbol{\xi}^{\mu(t)}, \boldsymbol{T}\cdot\boldsymbol{\xi}^{\mu(t)})\,\mathrm{d}t + \mathrm{d}\boldsymbol{W}(t). \tag{4.76}$$

The random Wiener process has components $\mathrm{d}W_n(t)$ which are normally distributed with mean 0 and variance $2D\,\mathrm{d}t = 2\,\mathrm{d}t$ in our choice of units. It is related to the noise term of the Langevin equation via $\mathrm{d}W_n(t) = \int_t^{t+\mathrm{d}t} \mathrm{d}t' \zeta_n(t')$; see [74] for more details. All other symbols take the same meaning as discussed before Eq. (4.8). We assume that the inputs that enter the equation are changing on a timescale much faster than the relaxation time of the weights. Hence it is only

the statistical properties of the inputs that determine the dynamics of $\boldsymbol{w}$ in the thermodynamic limit. We can thus average over the inputs, making the detailed dynamics of $\mu(t)$ unimportant. We can derive the equations of motion for the means of $\mathcal{Q} \equiv \boldsymbol{w} \cdot \boldsymbol{w}/N$ and $\mathcal{R} \equiv \boldsymbol{T} \cdot \boldsymbol{w}/N$ by expanding to *second* order in $\mathrm{d}\boldsymbol{w}$ and keeping terms on the order of $\mathrm{d}t$:

$$\begin{aligned}
\mathrm{d}\mathcal{Q} \equiv& \mathcal{Q}(\boldsymbol{w} + \mathrm{d}\boldsymbol{w}) - \mathcal{Q}(\boldsymbol{w}) \\
&= \frac{1}{N}\left(2\boldsymbol{w} \cdot \mathrm{d}\boldsymbol{w} + \mathrm{d}\boldsymbol{w} \cdot \mathrm{d}\boldsymbol{w}\right) \\
&= 2(1 - k\mathcal{Q})\,\mathrm{d}t + \tilde{\nu}(t)^2 \left\langle \mathcal{F}(\sqrt{\mathcal{Q}}, \boldsymbol{T} \cdot \boldsymbol{\xi}, \boldsymbol{w} \cdot \boldsymbol{\xi})^2 \right\rangle \mathrm{d}t \\
&\qquad + 2\tilde{\nu}(t) \left\langle \operatorname{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi}) \frac{\boldsymbol{w} \cdot \boldsymbol{\xi}}{\sqrt{N}} \mathcal{F}(\sqrt{\mathcal{Q}}, \boldsymbol{T} \cdot \boldsymbol{\xi}, \boldsymbol{w} \cdot \boldsymbol{\xi}) \right\rangle \mathrm{d}t
\end{aligned} \tag{4.77}$$

where, contrary to ordinary calculus, the term $\mathrm{d}\boldsymbol{w} \cdot \mathrm{d}\boldsymbol{w}$ has contributed two terms, one from the Wiener process and one because $\boldsymbol{\xi} \cdot \boldsymbol{\xi} \approx N \approx 1/\,\mathrm{d}t$. We have also used the scaling $\nu(t) = \sqrt{N}\tilde{\nu}(t)$ that we discussed in Sec. 4.6.1 of the main text. In this section, we will denote by $\langle \cdot \rangle$ the average with respect to the distribution of inputs (4.1). Likewise, we have

$$\begin{aligned}
\mathrm{d}\mathcal{R} &= \boldsymbol{T} \cdot \mathrm{d}\boldsymbol{w}/N \\
&= -k\mathcal{R}\,\mathrm{d}t + \tilde{\nu}(t) \operatorname{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi}) \frac{\boldsymbol{T} \cdot \boldsymbol{\xi}}{\sqrt{N}} \mathcal{F}\,\mathrm{d}t
\end{aligned} \tag{4.78}$$

We can simplify the averages over the inputs and thus these equations by noting that the inputs only enter as products

$$x \equiv \frac{\boldsymbol{T} \cdot \boldsymbol{\xi}}{\sqrt{N}} \quad \text{and} \quad y \equiv \frac{\boldsymbol{w} \cdot \boldsymbol{\xi}}{\sqrt{N}}, \tag{4.79}$$

such that

$$\mathrm{d}\mathcal{Q} = 2(1 - k\mathcal{Q})\,\mathrm{d}t + 2\tilde{\nu}(t) \left\langle \operatorname{sgn}(x) y \mathcal{F}(x, y) \right\rangle \mathrm{d}t + \tilde{\nu}(t)^2 \left\langle \mathcal{F}(x, y)^2 \right\rangle \mathrm{d}t \tag{4.80}$$

$$\mathrm{d}\mathcal{R} = -k\mathcal{R}\,\mathrm{d}t + \tilde{\nu}(t) \left\langle \operatorname{sgn}(x) x \mathcal{F}(x, y) \right\rangle \mathrm{d}t \tag{4.81}$$

The crucial point to compute the averages $\langle \cdot \rangle$ is now to realise that $p(x, y)$ is a binormal Gaussian distribution due to the central limit theorem, with moments

$$\begin{aligned}
&\langle x \rangle = \langle y \rangle = 0 \\
&\left\langle x^2 \right\rangle = 1, \left\langle y^2 \right\rangle = \mathcal{Q} \\
&\langle xy \rangle = \frac{1}{N} \sum_{n,m} T_m w_n \left\langle \xi_n \xi_m \right\rangle = \frac{\boldsymbol{T} \cdot \boldsymbol{w}}{N} = \mathcal{R}
\end{aligned}$$

where we have used $\langle \xi_n \rangle = 0$ and $\langle \xi_n \xi_m \rangle = \delta_{nm}$ from Eq. (4.1). Here we only give the results of these integrals for the different learning algorithms for completeness, see *e.g.* [113] for details on how to perform these integrals*.

For Hebbian learning, we have $\mathcal{F} = 1$ and hence

$$\langle \mathrm{sgn}(x) y \rangle = \frac{2}{\pi} \mathcal{R}, \tag{4.82}$$

$$\langle \mathrm{sgn}(x) x \rangle = \frac{2}{\pi}. \tag{4.83}$$

The perceptron algorithm has $\mathcal{F}(x, y) = \theta(-xy)$ such that

$$\langle \mathrm{sgn}(x) y \mathcal{F} \rangle = \frac{\mathcal{R} - \sqrt{\mathcal{Q}}}{\sqrt{2\pi}}, \tag{4.84}$$

$$\langle \mathrm{sgn}(x) x \mathcal{F} \rangle = \frac{1 - \mathcal{R}/\sqrt{\mathcal{Q}}}{\sqrt{2\pi}}, \tag{4.85}$$

$$\left\langle \mathcal{F}^2 \right\rangle = \frac{1}{2} - \frac{1}{\pi} \operatorname{atan} \frac{\mathcal{R}}{\sqrt{\mathcal{Q} - \mathcal{R}^2}}. \tag{4.86}$$

Finally, for AdaTron learning with $\mathcal{F}(x, y) = |y| \theta(-xy)$, we find

$$\langle \mathrm{sgn}(x) y \mathcal{F} \rangle = \frac{\mathcal{R}\sqrt{\mathcal{Q} - \mathcal{R}^2}}{\pi} + \mathcal{Q} \left(-\frac{1}{2} + \frac{1}{\pi} \operatorname{atan} \frac{\mathcal{R}}{\sqrt{\mathcal{Q} - \mathcal{R}^2}} \right), \tag{4.87}$$

$$\langle \mathrm{sgn}(x) x \mathcal{F} \rangle = \frac{\sqrt{\mathcal{Q} - \mathcal{R}^2}}{\pi} + \mathcal{R} \left(-\frac{1}{2} + \frac{1}{\pi} \operatorname{atan} \frac{\mathcal{R}}{\sqrt{\mathcal{Q} - \mathcal{R}^2}} \right), \tag{4.88}$$

$$\left\langle \mathcal{F}^2 \right\rangle = -\frac{\mathcal{R}\sqrt{\mathcal{Q} - \mathcal{R}^2}}{\pi} + \frac{\mathcal{Q}}{2\pi} \left(\pi - 2 \operatorname{atan} \frac{\mathcal{R}}{\sqrt{\mathcal{Q} - \mathcal{R}^2}} \right). \tag{4.89}$$

Substituting these results into the equations for $\dot{\mathcal{Q}}$ and $\dot{\mathcal{R}}$, Eq. (4.37), yields a closed set of equations in $\mathcal{Q}$ and $\mathcal{R}$, which can be solved numerically.

4.F Computing the excess heat

The most straightforward way to compute the excess heat for learning in the thermodynamic limit $N \to \infty$ after a quench of the learning rate to $\tilde{\nu}(t) = \tilde{\nu}_0$ is by relying on the machinery developed in Section 4.E, namely the Itô stochastic differential equation for the weights, Eq. (4.76), which we rewrite slightly here as

$$\mathrm{d}\boldsymbol{w} = F(\boldsymbol{T}, \boldsymbol{w}, \boldsymbol{\xi}) \, \mathrm{d}t + \mathrm{d}\mathbf{W}(t) \tag{4.90}$$

* N.B. they use a different normalisation procedure from us.

with the total force on the weights $F(\boldsymbol{T}, \boldsymbol{w}, \boldsymbol{\xi})$. The key insight here is due to K. Sekimoto, who realised that this equation is indeed a statement of the first law, with the stochastic heat increment defined as [64, 78]

$$đq \equiv F(\boldsymbol{T}, \boldsymbol{w}, \boldsymbol{\xi}) \circ \mathrm{d}\boldsymbol{w} = \frac{1}{2}\left(F(\boldsymbol{T}, \boldsymbol{w}, \boldsymbol{\xi}) + F(\boldsymbol{T}, \boldsymbol{w} + \mathrm{d}\boldsymbol{w}, \boldsymbol{\xi})\right) \mathrm{d}\boldsymbol{w} \tag{4.91}$$

where we have evaluated the stochastic product $\circ$ using the Stratonovich or mid-point convention for every component [74]. For the excess heat, we replace the total force on the weights with the gradient of the "non-equilibrium" potential [20]

$$\phi(\boldsymbol{T}, \boldsymbol{w}; \tilde{\nu}_0) \equiv -\ln p^{\mathrm{s}}(\boldsymbol{T}, \boldsymbol{w}; \tilde{\nu}_0) \tag{4.92}$$

where $p^{\mathrm{s}}(\boldsymbol{T}, \boldsymbol{w}; \tilde{\nu}_0)$ is the steady-state distribution for $\tilde{\nu}_0$ with steady-state values $\mathcal{Q}^{\mathrm{s}}(\tilde{\nu}_0)$ and $\mathcal{R}^{\mathrm{s}}(\tilde{\nu}_0)$. Hence after averaging over the thermal noise, we find for the average increment in excess heat of the system over N

$$đQ_n^{\mathrm{ex}} = \frac{1}{N} đQ^{\mathrm{ex}} \equiv -\frac{1}{N}\, \mathrm{d}\boldsymbol{w} \circ \nabla_{\boldsymbol{w}} \phi(\boldsymbol{T}, \boldsymbol{w}; \tilde{\nu}_0) \tag{4.93}$$

$$= -\frac{1}{N}\, \mathrm{d}\boldsymbol{w} \circ \frac{\boldsymbol{w} - \mathcal{R}^{\mathrm{s}}(\tilde{\nu}_0)\boldsymbol{T}}{\mathcal{Q}^{\mathrm{s}}(\tilde{\nu}_0) - \mathcal{R}^{\mathrm{s}}(\tilde{\nu}_0)^2} \tag{4.94}$$

$$= \frac{1}{\mathcal{Q}^{\mathrm{s}}(\tilde{\nu}_0) - \mathcal{R}^{\mathrm{s}}(\tilde{\nu}_0)^2} [k(\mathcal{Q}(t) - \mathcal{R}^{\mathrm{s}}(\tilde{\nu}_0)\mathcal{R}(t)) - \tilde{\nu}_0 \left\langle \mathrm{sgn}(x) y \mathcal{F}(x, y) \right\rangle$$

$$+ \tilde{\nu}_0 \mathcal{R}^{\mathrm{s}}(\tilde{\nu}_0) \left\langle \mathrm{sgn}(x) x \mathcal{F}(x, y) \right\rangle - \tilde{\nu}_0^2 \left\langle \mathcal{F}(x, y)^2 \right\rangle / 2 - 1]\, \mathrm{d}t. \tag{4.95}$$

where $\langle \cdot \rangle$ now indicates an average over the distribution $p(x, y)$ as described in Appendix 4.E and we remind ourselves that we chose units where $D = 1$.

5 Universal costs of learning and the time-energy-information trade-off

We have analysed a number of different learning problems and different learning algorithms in the previous chapters, *i.e.* we studied the learning of the labels of a fixed set of inputs as well as the learning of a rule in Chapters 3 and 4, and we analysed the Hebbian, Perceptron and AdaTron algorithms, respectively. We generally found that the thermodynamic costs of learning were an upper bound on the amount of information that a student could extract from data or learn from a teacher and that these inequalities hold for all times. These observations naturally lead to two questions.

The first question concerns time. Although our results hold at all times, it is intriguing to wonder about the explicit role of time for learning. Consider for example the fact that when learning a rule, the generalisation error saturates at a constant value after a certain learning duration τ, see for example Fig. 4.1. However, as we have seen in Section 4.5, any non-trivial neural network will relax to a non-equilibrium steady state with a constant rate of dissipation. This motivates the search for a trade-off between dissipation, information and learning duration. Such relationships have been found for for specific models of chemical sensing [9, 60] and for a general model of a physical sensor [11] in the limit where the sensor is evolving much faster than the signal is changing, *i.e.* for a sensor whose thermodynamics are described using linear response theory [14, 15]. This raises the question of whether such a trade-off can be derived for learning.

Second, one may of course ask whether the relationship between entropy production and learnt information that we found is universal for a reasonable class of learning problems and algorithms. What about unsupervised learning, where the data do not carry a label and the task is instead to discover some structure in the data, as for example in clustering? Another important extension of our work concerns more complicated neural architectures. The single-layer perceptrons we have studied so far are limited to the implementation of linearly separable Boolean functions $\sigma_{\mathrm{T}}(\boldsymbol{\xi}) = \pm 1$, such as the one shown in Fig. 5.1b. A natural generalisation is to add intermediate layers, where the output of the neurons in one layer is the input for the neurons in the next [8, 92]. An example of such a network, a soft committee machine with a single hidden layer, is shown schematically in

Fig. 5.1c and an example of a function that this network can implement in two dimensions is plotted in 5.1d. Here, we will not discuss multilayer networks in details and refer to the existing literature for a more systematic treatment [8, 92, 124] (but see Appendix 5.C for details about the neural network behind Fig. 5.1d). Instead, we will consider a general class of inference problems as described below that covers both deep networks, unsupervised learning etc. in a unified approach.

In this chapter, we answer both these questions in the affirmative. We prove an integral fluctuation theorem (IFT) for a very broad class of learning problems and algorithms using stochastic thermodynamics on Bayesian networks [125]. The

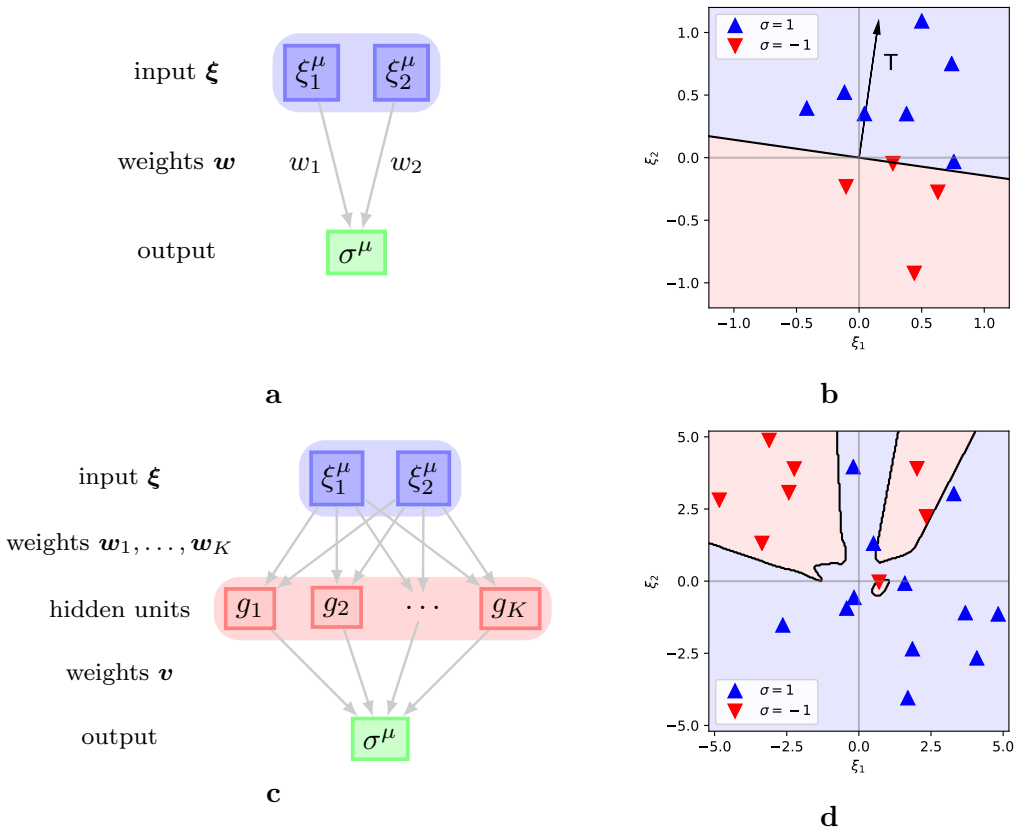


Figure 5.1 | Single- and two-layer neural networks and the functions they can implement. The top row is a reminder of the architecture (**a**) and an example of a function $\sigma_{\mathrm{T}}^{\mu} = \mathrm{sgn}(\boldsymbol{T} \cdot \boldsymbol{\xi}^{\mu})$ that the single-layer linear perceptron can implement (**b**). The function shown below (**d**) was implemented using the two-layer neural network (**c**), where the output $\sigma_{\mathrm{T}}^{\mu} = \mathrm{sgn}(\sum_{k=1}^{K} [v_k \, g(\boldsymbol{w}_k \cdot \boldsymbol{\xi}^{\mu} + b_k) + c])$ is the weighted sum of the state g_k of the hidden units. We chose $g(x) = \tanh(x)$, $K = 10$ (see Appendix 5.C for further details). Clearly, adding just a single layer of hidden units greatly increases the complexity of the functions that can be implemented (see also our discussion in Sec. 4.7).

second law that follows from this IFT shows that the entropy production of the weights is a universal bound for the information acquired during learning. Second, we will use a form of the thermodynamic uncertainty relation to derive a general relationship between energy, time and information that applies to learning.

5.1 Setup

Consider the following learning problem: a student observes a stream or a set of samples y, each a vector of arbitrary dimension drawn independently from a probability distribution $q(y|B)$. The parameters B may or may not change over time, but the functional form of q remains fixed. Both the functional form of the distribution $q(y|B)$ and the values of the parameters B are unknown to the student. To model the data, the student chooses a statistical model $p(y|w)$ which is parametrised by another set of real numbers w. The vectors B and w could for example represent the weights of a multi-layer neural network [8, 92] and they need not be of the same dimension. If $p(\cdot)$ and $q(\cdot)$ have the same functional form, the problem is *realisable*. The student's task is to reconstruct the unknown distribution q by inferring the optimal parameters $\hat{w}$ from the available data.

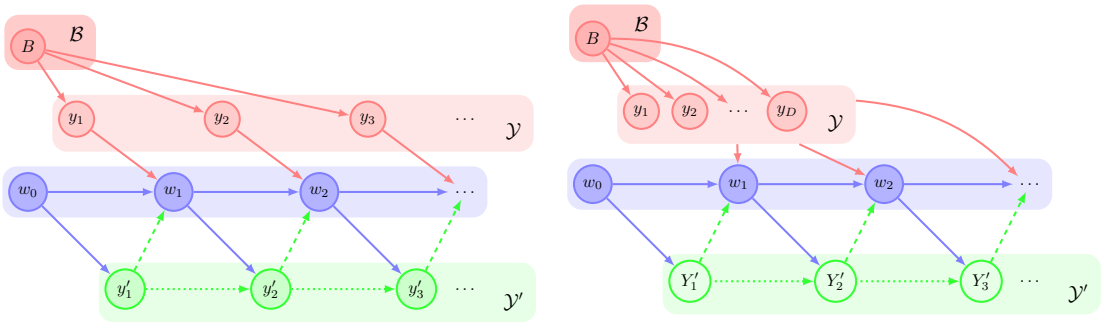
This setup naturally applies to unsupervised learning, where each y is typically a vector in a space of features*. Unsupervised learning is very important because most of the data in the world, including our sensory inputs, is unlabelled, and it plays a fundamental role in modern deep learning architectures [124, 127]. The setup also applies to supervised learning, which is indeed a special case of unsupervised learning†. In a supervised learning problem, the parameters B specify a function for the true label $\lambda = f_B(\xi)$ of inputs ξ. λ may be discrete for classification or continuous in regression problems, respectively. The student aims to infer the function $\xi \to \lambda$ from samples $y = (\lambda, \xi)$, such that the student can later predict the label λ for a previously unseen input ξ faithfully. In any realistic problem, the deterministic function $f_B(\xi)$ should be replaced with a data generating distribution $q(y|B) = q(\lambda|B, \xi)q(\xi)$ where $q(\lambda|B, \xi)$ is the probabilistic rule to be learned and the input distribution $q(\xi)$ is given.

5.2 Learning dynamics and Bayesian networks

We assume that the weights are evolving in contact with a heat bath at inverse temperature $\beta = 1$. Starting from an *a priori* set of parameters w_0 before

* A classic example is learning the symmetry-breaking direction B of the distribution of y [126]

† Consider for example the supervised learning of a rule $\boldsymbol{\xi} \to \sigma_{\mathrm{T}}$ that we discussed in Chapter 4. Learning this rule can be understood as the unsupervised learning of the vectors $y = (\sigma_{\mathrm{T}}, \xi)$ with a single symmetry-breaking direction, namely the first component of y.



a Online learning from a static teacher B. **b** Batch learning from a static teacher B.

Figure 5.1 | Bayesian networks for online and batch learning. These networks model the causal structure of the learning dynamics for the different learning problems. Each node in a network represents a random variable in a time series of the degrees of freedom. A directed edge of any kind (solid, dashed, etc.) indicates a causal relationship between the connected nodes. The degrees of freedom in our model are the parameters of the unknown distribution B, the samples y, the parameters of our model w and the outputs of our model, y'. The dashed lines only apply to learning with feedback, while the dotted lines only apply if the output process is a Markov chain itself.

any data is observed, the weights evolve as a discrete-time Markov process [74] with time $t = 1, 2, \ldots$*. The resulting trajectory or time series of the weights $w_{0:\tau} \equiv \{w_t \,|\, t = 0, 1, \ldots, \tau\}$ is sketched in the graphs of Fig. 5.1 for the – broadly speaking – two different approaches to learning from examples that exist, batch and online learning.

The learning dynamics are specified by transition probabilities, *e.g.* $p(w_t \to w_{t+1})$ for the weights or $p(y'_t \to y'_{t+1})$ for the output. Their exact form depends on the learning algorithm chosen and is not important for our results. What is important is which variables contribute to the update of the weights at every time step. During *online learning*, shown in Fig. 5.1a, a single sample y_t is drawn from the distribution $q(y|B)$ at every time step, used to update the weights and then discarded. Online learning hence requires little memory and is well suited for learning problems with time-dependent parameters B, see below. *Batch learning*, sketched in Fig. 5.1b, requires a data set $\mathcal{Y} = \{y_1, y_2, \ldots, y_D\}$ where each sample y_n is drawn independently from $q(y|B)$ at $t = 0$. The set $\mathcal{Y}$ is available at every

* N.B. that the IFT (5.8) and hence Eq. (5.9) also apply to models with continuous time $w(t)$ which can be discretised by introducing an infinitesimal time interval Δt and splitting the trajectory into $\tau/\Delta t \gg 1$ intervals. Each interval is labelled by t_n with $w_n = w(n\Delta t)$. The resulting time series can be modelled as described in the main text. Care has to be taken in the handling of noise and the Stratonovich convention should be used [20, 74]; see for example [128] for a technical discussion

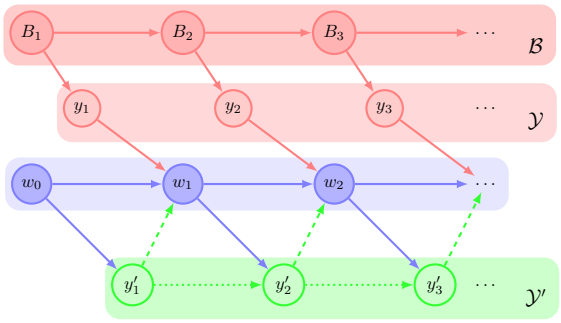


Figure 5.2 | Bayesian network for online learning from a time-dependent teacher B.

update of the weights*.

Some learning algorithms additionally apply feedback to the weights w_t using the current output of the model, indicated by the dashed lines in Fig. 5.1. For online learning, this is just a single output of the model, y', while for batch learning, this can involve the output of the model corresponding to all the data points in the data set $\mathcal{Y}$, which we will denote by Y'. Classic examples of learning algorithms with feedback are backpropagation [129] for supervised learning in multi-layer neural networks and contrastive divergence [130, 131] in restricted Boltzmann machines for unsupervised learning. Some algorithms furthermore require that the outputs of the model themselves are a Markov model and not independently drawn† (dotted lines in Fig. 5.1).

Learning something once is usually not enough. Instead, a learner has to adapt to a changing environment, which we model as a time-dependent set of parameters B, for example for a student learning from a time-dependent teacher [120, 121, 133]. We model such a teacher as a Markov chain $B_1, B_2, \ldots$, where y_t is drawn from the distribution $q(y_t|B_t)$, and so on. In such a scenario, online learning is the most sensible approach. The resulting Bayesian network for the dynamics is shown in Fig. 5.2.

The update rules for the weights are hence given by conditional probabilities, for example $p(w_t|w_{t-1}, y_t, y'_t)$ for an online algorithm (only one sample y_t enters) with feedback (the update depends on y'_t). A batch algorithm without feedback has $p(w_t|w_{t-1}, \mathcal{Y})$, where the full data set $\mathcal{Y}$ enters, but not any outputs.

Learning involves three distinct sources of randomness: The heat bath causes thermal fluctuations in the weights. There is randomness in the data, either due to the random succession of samples y_t for online learning or due to the finite size of the data set $\mathcal{Y}$ for batch learning. Finally, there is either quenched disorder in

* There exist also intermediate forms between batch and online learning, where a large training set is split into a number of smaller sets or *mini-batches*.

† An example is persistent contrastive divergence for restricted Boltzmann machines [132].

the form of the parameters B in case these parameters remain fixed, or additional noise if the dynamics of B are stochastic.

5.2.1 Bayesian networks

The graphs in Figs. 5.1 and 5.2 are not just illustrations of the learning dynamics, but can instead be interpreted as Bayesian networks [125] which define the stochastic learning dynamics in the following way. A Bayesian network consists of a set of M nodes $\mathcal{X} = \{x_1, x_2, \ldots, x_M\}$ representing stochastic variables and a set of edges connecting them. The state of every system at every time step corresponds to one node, *e.g.* w_1. An edge $x_i \to x_j$ between nodes indicates a causal relationship between these nodes in the direction of the edge and we say that x_i is a parent of x_j. The parents of a node are denoted by $\mathrm{pa}(x_m)$, *e.g.* $\mathrm{pa}(w_2) = \{w_1, y_2', y_2\}$ in Fig. 5.1a. The key property of Bayesian networks is that the nodes are in topological ordering, *i.e.* the order is determined by the causal relationship between the nodes in the sense that x_j cannot be the parent of x_i if $i > j$.

We get the full probability distribution of the time evolution of the whole system by the chain rule of probability and the identity $p(x_n|x_{n-1}, \ldots, x_0) = p\left(x_n|\,\mathrm{pa}(x_n)\right)$ (see Appendix 5.A), such that $p(\mathcal{X}) = \prod_{m=1}^{M} p(x_m|\,\mathrm{pa}(x_m))$. The ensemble average of an arbitrary function $f(\mathcal{X})$ is given by $\langle f \rangle = \sum_{\mathcal{X}} p(\mathcal{X}) f(\mathcal{X})$.

5.2.2 Entropy production and information processing

Stochastic thermodynamics on Bayesian networks was first formulated by Ito and Sagawa [134]. The key quantity of this approach is still the total entropy production of the weights along a single trajectory σ_w [82]. The increase in entropy of the surrounding heat bath due to the dissipation of heat [20] Δq_w is given by

$$\Delta q_w \equiv \sum_{t=0}^{\tau-1} \ln \frac{p(w_{t+1}|w_t, \mathcal{P}_{t+1})}{p^\dagger(w_t|w_{t+1}, \mathcal{P}_{t+1})} \tag{5.1}$$

where $p^\dagger$ is the probability distribution of the backward step and

$$\mathcal{P}_{t+1} \equiv \mathrm{pa}(w_{t+1}) \setminus\ w_t, \tag{5.2}$$

with $\setminus$ indicating the relative complement of two sets. The resulting change in the entropy of the medium is defined as $\Delta s_w^{\mathrm{m}} \equiv \beta \Delta q_w$. The change in stochastic entropy of the system remains $s(t) \equiv -\ln p(w_t)$, which is the logarithm of the ensemble probability of the system evaluated with the weights of time τ [82], see

also Section 2.1.2. The apparent entropy production of the weights along a single trajectory thus reads

$$\sigma_w \equiv \ln p(w_0) - \ln p(w_\tau) + \Delta s^{\mathrm{m}}_w \tag{5.3}$$

We have introduced the mutual information as a key concept for the analysis of information processing. Here, we will study it on a trajectory level, so we define for the conditional mutual information [31], which measures the correlations between the random variables A and B given knowledge of another random variable C:

$$I(A:B\,|\,C) \equiv \left\langle \ln \frac{p(a,b|c)}{p(a|c)p(b|c)} \right\rangle \equiv \langle i(A:B|C)\rangle \tag{5.4}$$

where the average is taken over $p(a,b,c)$. Note that each of A, B and C can stand for a number of random variables, for example a whole trajectory $w_{0:t}$. For brevity, we will write $i(A:B|\varnothing) = i(A:B) \equiv \ln p(A,B) - \ln p(A) - \ln p(B)$.

Finally, we introduce the transfer entropy, an information theoretic measure of correlations between two time series $a_{0:\tau}$ and $b_{0:\tau}$ [135]. It quantifies by how much the uncertainty about the next state of the process y, given the history of y, is reduced by knowing the history of the x process:

$$T^{a\to b}_\tau \equiv I(b_\tau : b_{0:\tau-1}, a_{0:\tau-1}) - I(b_\tau : b_{0:\tau-1}) \tag{5.5}$$

$$= I(b_\tau : a_{0:\tau-1}\,|\,b_{0:\tau-1}) \geq 0. \tag{5.6}$$

The last inequality follows from the positivity of mutual information. The transfer entropy explicitly takes into account the *direction* of the information flow such that $T^{a\to b} \neq T^{b\to a}$, contrary to the mutual information, which is symmetric in its arguments: $I(a:b) = I(b:a)$. The rate of transfer entropy is given by

$$\dot{T}^{a\to b}_t = I(b_t : a_{t-1}\,|\,b_{0:t-1}) \tag{5.7}$$

5.3 Universal costs of learning

We are now in a position to formulate our main result. Let us denote by $\mathcal{Y}$ the set of examples that the learner has seen up to a time τ and by $\mathcal{Y}'$ the set of all the outputs of the neuron up to time τ, *i.e.* $\mathcal{Y} = \{y_n | n = 1, \dots N\}$, $\mathcal{Y}' = Y'_{1:\tau}$ for batch and $\mathcal{Y} = y_{1:\tau}$, $\mathcal{Y}' = y'_{1:\tau}$ for online learning, respectively. Similarly, we define $\mathcal{B} = B_{1:\tau}$ for time-dependent parameters B or just $\mathcal{B} = B$ for a static $q(y|B)$.

We can then show that for any learning algorithm that gives rise to a Bayesian network of the forms shown in Figs. 5.1 and 5.2, the following integral fluctuation theorem (IFT) holds for all learning durations t^*:

$$\langle \exp\left(-\sigma_w + i(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}')\right)\rangle = 1, \tag{5.8}$$

* The derivation is given in 5.B

Using Jensen's inequality for convex functions as $\langle \exp(x) \rangle \geq \exp(\langle x \rangle)$ [8], we have

$$\langle \sigma_w \rangle \geq I(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}') \geq I(w_\tau : \mathcal{B}) \geq 0, \tag{5.9}$$

where we used the chain rule of mutual information for the second inequality [31]. This is the second law of thermodynamics for learning from examples. We emphasise that our result applies to regular and non-regular learning problems and does not require the limits $N \to \infty$ or $t \to \infty$.

How sharp is the bound $I(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}') \geq I(w_\tau : \mathcal{B})$? In the limit of large data sets or long times, the posterior distribution of the weights will be strongly concentrated around its maximum $\hat{w}$. If we assume that $p(y'|w)$ is a smooth function of its parameters, we can approximate the asymptotic posterior distribution by a Gaussian distribution of the form

$$p(w|\mathcal{Y}) \sim \exp\left(-\frac{1}{2}(w - \hat{w})J^{-1}(w - \hat{w})\right) \tag{5.10}$$

where the matrix $J_{ij} = -\partial_i \partial_j \sum_n \ln p(y'_n|\hat{w})$ is the Fisher information [31] with $\partial_i \equiv \partial/\partial \hat{w}_i$. This expansion can be justified rigorously if the problem is realisable [136] and is determined by a set of just $\mathcal{O}(M^2)$ quantities, where M is the number of parameters in w. Hence all the information in the training set $\mathcal{Y}$ is asymptotically compressed into a much smaller set of quantities which determine the posterior distribution of w. We can therefore expect that the inequality is increasingly sharp as the learning continues.

5.3.1 Feedback and transfer entropy

It is also instructive to apply the IFT by Ito and Sagawa for causal networks to our setup [134]. Here, we will again focus on a static teacher; the generalisation to a time-dependent teacher is straightforward and discussed below. Applying the integral fluctuation theorem for Bayesian networks [134] to the dynamics of learning, we find that for all learning durations τ

$$\langle \exp(-\sigma_w + \Theta^{\mathrm{x}}) \rangle = 1, \tag{5.11}$$

where Θ is a quantity that captures the information flows across the network and $x = o, b$ for online and batch learning, respectively. Specifically, we have for *online learning* (Fig. 5.1 a)

$$\Theta^{\mathrm{o}} = i(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}') - \sum_{t=1}^{\tau} i(y'_t : w_{t-1}|B_{1:t}, y_{1:t}, y'_{1:t-1}) \tag{5.12}$$

where we can identify the last term as the sum over the transfer entropy rates from the weights into the other systems B, y, and y'. Using Jensen's inequality for convex functions as $\langle \exp(x) \rangle \geq \exp(\langle x \rangle)$ [31], we can derive a second-law like statement from Eq. (5.11),

$$\langle \sigma_w \rangle + \sum_{t=1}^{\tau} I(y'_t : w_{t-1} \,|\, B, y_{1:t}, y'_{1:t-1}) \geq I(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}') \geq I(w_\tau : B), \tag{5.13}$$

where we used the chain rule of mutual information for the last inequality [31]. We emphasise that while the second IFT (5.11) is not a straightforward extension of our main result, Eq. (5.8), the second-law like inequality found by including the transfer entropy (5.13) is of course weaker than our previous bound, Eq. (5.9), since we can always add a positive quantity to the latter, for example a transfer entropy.

For batch learning (Fig. 5.1 b), we define $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ and the output of the model at step t, $Y'_t \equiv \{y'_1, y'_2 \ldots, y_N\}$, and find

$$\Theta^{\mathrm{b}} = i(w_\tau : B, \mathcal{Y}, Y'_{1:\tau}) - \sum_{t=1}^{\tau} i(Y'_t : w_{t-1} | B, \mathcal{Y}, Y'_{1:t-1}), \tag{5.14}$$

with the corresponding second-law.

5.3.2 Refining the IFT

In many learning problems, for example online learning, the weights relax to a non-equilibrium steady state with a constant rate of heat dissipation due to the driving from the fluctuating stream of changing samples y. While the IFT (5.8) still holds, the second law (5.9) is not very sharp anymore: the total entropy production increases with time at a constant rate, while the mutual information remains on the order of one. This observation motivates the following refinement of the IFT using concepts from steady state thermodynamics (SST) [86, 87, 137], which we discussed in Sections 2.5 and 4.5, so here we just briefly remind ourselves that the key idea of SST is to split the entropy production σ_w into two contributions

$$\sigma_w = \sigma_w^{\mathrm{na}} + \sigma_w^{\mathrm{a}}. \tag{5.15}$$

which each correspond to a mechanism that drives a system out of equilibrium: applying non-equilibrium constraints such as a fluctuating force leads to adiabatic entropy production σ_w^{a} to maintain the steady state. Driving the system from one steady state to another will additionally lead to non-adiabatic entropy production σ_w^{na}, which enters our refined IFT (see Sec. 5.B for details)

$$\langle \exp\left(-\sigma_w^{\mathrm{na}} + i(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}')\right) \rangle = 1. \tag{5.16}$$

The refined second law $\langle \sigma_w^{\mathrm{na}} \rangle \geq I(w_\tau : \mathcal{B})$ is a much sharper inequality than Eq. (5.9) since the average rate of non-adiabatic entropy production becomes identically 0 in the steady state, such that σ_w^{na} is on the order of 1.

5.3.3 Generalising from the data

The mutual information $I(w_\tau : B)$ is a thermodynamically consistent measure of learning by virtue of Eq. (5.8). It measures the quality of the weights w as estimators of the parameters B. However, we are often interested in the mutual information of the output of the model y and that of the student y', *e.g.* for supervised learning, where the quantity of interest is the mutual information between the true label λ and the student's prediction λ'. The connection between the two can be made by using the data processing inequality $I(X : Y) \geq I(X : Z)$ for the Markov chain $X \to Y \to Z$ [31] to obtain

$$\langle \sigma_w \rangle \geq I(w_\tau : B) \geq I(\lambda : \lambda'). \tag{5.17}$$

The ability to generalise is hence also related to the total entropy production of the weights during learning.

5.4 Time-energy-information trade-off for learning

The IFT (5.8) relates the dissipated free energy to the information that is acquired during learning. While the IFT holds for all learning durations t, time does not enter the expression explicitly. We now derive a general trade-off where the learning duration t enters explicitly and which is based on the thermodynamic uncertainty relation. This relation was discovered by Barato and Seifert, who provided a proof in the linear response regime and for unicyclic networks arbitrarily far from equilibrium [63]. It was conjectured as a universal bound on current fluctuations based on arguments from large deviation theory and extensive numerics [138] and proven shortly thereafter [139]. The finite time uncertainty relation that we will use here was conjectured in Ref. [140] and proven in Ref. [141].

For our purposes, we consider the continuous-time Markovian dynamics of the system (B_t, y_t, w_t, y'_t). For concreteness, we will assume that the state space of this system is discrete* and has N states $\{n\}$ in total. Transitions from state n to state m take place with rate $k_{nm} \geq 0$. All rates are taken to be reversible, *i.e.* $k_{nm} \geq 0 \Leftrightarrow k_{mn} \geq 0$, for thermodynamic consistency [20]. We will further impose

* N.B. Since the thermodynamic uncertainty relation also holds for diffusive processes [142], the results of this chapter also apply for continuous systems obeying Fokker-Planck type dynamics in continuous time, such as the models of Chapter 3 and 4

the restriction on the rates that the system is *multipartite* [38, 84]: hence, in every allowed transition, only *one* of the degrees is changed, *i.e.* either B changes, while y, w and y' stay constant, *etc.* This equivalent to the assumption that we have made throughout this thesis that the thermal noise sources in the individual degrees of freedom of a system are uncorrelated to each other; see Sec. 2.4 or the discussion on p. 65. The ensemble distribution $p(n,t)$ of the total system then obeys a master equation[73]

$$\partial_t p(n,t) = \sum_m \mathcal{L}_{nm}\, p(n,t) \equiv \sum_m \left[k_{mn} - r_n \delta_{nm}\right]\, p(n,t) \tag{5.18}$$

where we have introduced the exit rate $r_n \equiv \sum_m k_{nm}$. We will focus on the steady state, where the system has settled into its steady state distribution $p^{\mathrm{s}}(n)$ which obeys $\sum_m \mathcal{L}_{nm}\, p^{\mathrm{s}}(n) = 0$.

We are interested in the learning current $L(t)$ with $L(0) = 0$. This is a random variable, and for every transition $n \to m$ that involves any one of the weights, it is increased by a quantity

$$l_{nm} = \ln \frac{p(m)}{p(n)} = \ln \frac{p(w^{(m)}|B^{(n)}, y^{(n)}, y^{(n')})}{p(w^{(n)}|B^{(n)}, y^{(n)}, y^{(n')})} = -l_{mn} \tag{5.19}$$

where $w^{(m)}$ is the value of the weight in the state m and we have used the fact that the multipartite nature of our dynamics implies that $B^{(m)} = B^{(n)}$, *etc.*, for every transition that involves only a change of (any of) the weights.

It can thus be seen by inspection that the stochastic current $L(t)$ will on average increase with a rate that is given by the thermodynamic learning rate [54] of the weights l_w. However, being a stochastic variable, the variance of $L(t)$ will also increase with time. We thus define the *reliability of learning* as the inverse of the variance of the learning current:

$$\mathcal{R}^{-1} \equiv \mathrm{var}[L(t)] \tag{5.20}$$

Using the finite-time thermodynamic uncertainty relation [140, 141], we find a general inequality involving the learning duration t, the average of total entropy production of the system $\langle\sigma\rangle$, the reliability of learning that we just defined and the thermodynamic learning rate of the weights which reads

$$\langle\sigma\rangle \geq 2\, t\, \mathcal{R}\, l_w^2. \tag{5.21}$$

This result implies a general trade-off where an undesirable quantity, the average entropy production of the whole system, bounds the product of two desirable quantities: the thermodynamic learning rate l_w, which measures the speed of learning, and the reliability of learning $\mathcal{R}$. It is important to note that $\langle\sigma\rangle$ is the total

entropy production of the entire system. It is often reasonable to assume that the process that generates the examples is in equilibrium, so for a system that learns without feedback, the total entropy production becomes the entropy production of the degrees of freedom of the model. We will discuss some applications of the trade-off (5.21) in a forthcoming publication.

Let us finally note that a naive application of (5.21) to the (finite-time) thermodynamic uncertainty relation to discrete time processes is not possible [143]. However, a discrete-time formulation of the uncertainty relation was recently derived in the limit of infinitely long trajectories [144], which applies to the discrete time dynamics that underlie the Bayesian networks we discussed at the beginning of this chapter. It does not include time explicitly, but also gives a universal trade-off between dissipation and the speed and reliability of learning. This discrete-time uncertainty relation is particularly suited for the investigation of small learning systems, where there is no thermodynamic limit and it is hence not easily possible to approximate the discrete-time process with a process in continuous-time [113].

5.5 Concluding perspectives

This chapter in a sense completes the analysis of the thermodynamic costs of learning which was the goal we set for ourselves in the beginning. We have shown that the total entropy production of the degrees of freedom of an arbitrary probabilistic model is an upper bound on the information that a learner can extract from data or learn from a teacher for a very broad class of learning problems and algorithms. Our setup includes both supervised and unsupervised learning problems, and applies to any learning algorithm that produces the same causal structure for the learning dynamics as the ones shown in Figs. 5.1 and 5.2, which includes the learning problems analysed in Chapters 3 and 4.

We have also found a general inequality for learning that is based on the thermodynamic uncertainty relation and achieves two things. First, it makes the role of time explicit. Furthermore, the time-energy-information trade-off for learning shows that for a constant rate of dissipation, one can only improve the speed of learning at the expense of the reliability of learning, or vice versa.

The thermodynamic uncertainty relation is still a very young and promising field. The (finite-time) uncertainty relations and its "cousins" [145, 146] mostly involve the empirical currents and the exact probability densities. It would be intriguing to look for uncertainty relations which involve the empirical densities instead to obtain relations which are, in a sense, truly trajectory-based.

Appendices to chapter 5

The appendices to this chapter start by summarising a number of basic properties of Bayesian For a detailed treatment of Bayesian networks, see for example [125]. We then review the identification of heat and excess in stochastic thermodynamics for causal networks and proceed to prove the integral fluctuation theorems (5.8).

5.A Two basic properties of Bayesian networks

Here, we briefly state two basic properties of Bayesian networks that we will use in our derivation of the integral fluctuation theorem (5.8). We refer the interested to any of a number of textbooks on the topic for the proofs and further details on Bayesian networks, for example Ref. [125].

A Bayesian network consists of a set of random variables $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, each of which has a set of mutually exclusive states. The variables are connected by directed edges and form an acyclic directed graph*. The parents of node x_i are all variables x_j which have a directed edge $x_j \to x_i$ and are denoted by $\mathrm{pa}(x_i)$.

For a general set of random variables $\mathcal{X}$, the chain rule of probability states that

$$p(\mathcal{X}) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\cdots p(x_N|x_1, x_2, \ldots, x_{N-1}) \tag{5.22}$$

This is simplified in a Bayesian network by the following theorem:

Theorem 1. *The full probability distribution $p(\mathcal{X})$ is uniquely specified by a Bayesian network and can be written as*

$$p(\mathcal{X}) = \prod_{i=1}^{N} p(x_i|\,\mathrm{pa}\,x_i). \tag{5.23}$$

A second useful property is the consistency of the specification of a Bayesian network, provided by the following theorem:

Theorem 2. *If $\mathcal{X}'$ is a subset of $\{x_1, x_2, \ldots, x_{n-1}\}$ and $\mathrm{pa}(x_n)$ is a subset of $\mathcal{A}'$,* i.e. $\mathrm{pa}(x_n) \subset \mathcal{A}' \subset \{x_1, x_2, \ldots, x_{n-1}\}$*, then*

$$p(x_n|\mathcal{A}') = p\left(x_n|\,\mathrm{pa}(x_n)\right) \tag{5.24}$$

Proof. See the supplemental material of Ref. [134]. □

* Hence there is no directed path $x_1 \to \cdots \to x_n$ such that $x_1 = x_n$.

5.B Proof of the integral fluctuation theorems

The proof of the integral fluctuation theorem (5.8) proceeds in two steps. We first rewrite the trajectory-level mutual information using properties of Bayesian networks in Sec. 5.A,

$$i(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}') = \ln\left(\frac{p(w_\tau, \mathcal{B}, \mathcal{Y}, \mathcal{Y}')}{p(w_\tau)p(\mathcal{B}, \mathcal{Y}, \mathcal{Y}')}\right) \tag{5.25}$$

$$= \ln\left(\frac{p(w_\tau, \mathcal{B}, \mathcal{Y}, \mathcal{Y}')}{p(w_\tau)p(\mathcal{A})}\prod_{t=0}^{\tau} p(w_t|\,\mathrm{pa}(w_t))\right) \tag{5.26}$$

where all symbols take the meaning introduced in the main text and we have used $p(\mathcal{A}) = p(\mathcal{B}, \mathcal{Y}, \mathcal{Y}')\prod_{t=0}^{\tau} p(w_t|\,\mathrm{pa}(w_t))$. From this, we have

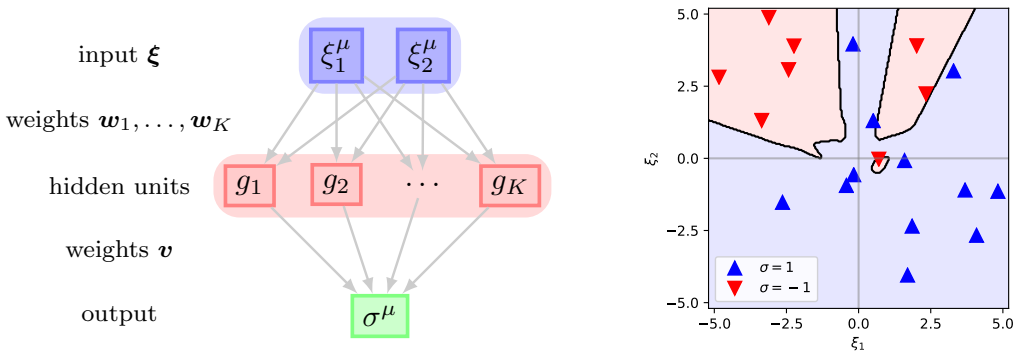
$$\left\langle \exp(-\Delta s_w^{\mathrm{tot}} + i(w_\tau : \mathcal{B}, \mathcal{Y}, \mathcal{Y}'))\right\rangle \tag{5.27}$$

$$= \sum_{\mathcal{A}} p(\mathcal{A})\frac{p(w_\tau)}{p(w_0)}\left(\prod_{t=0}^{\tau-1}\frac{p^\dagger(w_t|w_{t+1}, \mathcal{P}_{t+1})}{p(w_{t+1}|w_t, \mathcal{P}_{t+1})}\right)\frac{p(w_\tau, \mathcal{B}, \mathcal{Y}, \mathcal{Y}')}{p(w_\tau)p(\mathcal{A})}\prod_{t=0}^{\tau} p(w_t|\,\mathrm{pa}(w_t)) \tag{5.28}$$

$$= \sum_{\mathcal{A}} p(w_\tau, \mathcal{B}, \mathcal{Y}, \mathcal{Y}')\prod_{t=0}^{\tau-1} p^\dagger(w_t|w_{t+1}, \mathcal{P}_{t+1}) \tag{5.29}$$

$$= 1 \tag{5.30}$$

where we have used $\mathrm{pa}(w_{t+1}) \subseteq \{w_t, \mathcal{P}_{t+1}\}$. Here, we have generically written $p^\dagger(\cdot)$ for the probability of the backward dynamics, without specifying whether the backwards dynamics are reversed or dual reversed[20]. The calculation remains unchanged in both cases, since the key requirement is simply that $p^\dagger(\cdot)$ denotes the probability of the backwards process. By using the reversed dynamics, we recover the apparent entropy production of the weights involving the heat (5.1), while using the dual reversed dynamics yields the non-adiabatic entropy production involving the excess heat.



Two-layer neural networks and the functions they can implement. Figure reproduced from p. 94.

5.C Details on the neural network discussed in Fig. 5.1

Here we briefly describe the multi-layer neural network discussed in the introduction. We repeat the plots of its architecture and of an example function that this network can implement above for convenient reference.

This network has two layers of weights: the first (counting from the top) connects all the components of the inputs, here $\boldsymbol{\xi} \in \mathbb{R}^2$, to every *hidden unit*. At each of the K hidden units, an activation $g(\boldsymbol{w}_k \cdot \boldsymbol{\xi}) + b_k$ is computed, where $b_k \in \mathbb{R}$ is an offset independent of the inputs. The sign of the sum of the activations of the K hidden inputs, weighted by the second layer of weights $\boldsymbol{v} \in \mathbb{R}^K$, plus an offset c, is then the output, *i.e.*

$$\sigma_{\mathrm{T}}^\mu = \operatorname{sgn}\left(\sum_{k=1}^{K} \left[v_k \; g(\boldsymbol{w}_k \cdot \boldsymbol{\xi}^\mu) + b_k\right] + c\right) \tag{5.31}$$

We chose $K = 10$ and $g(x) = \tanh(x)$. This choice of a strongly non-linear function is key for the expressivity of deep networks, *i.e.* the large number of functions that can be implemented with these networks. The weights $\boldsymbol{w}_k$, $\boldsymbol{v}$ and the offsets b_k, c were drawn independently at random from a normal distribution with mean zero and standard deviation 5, 2, 4, 3, respectively.

This particular type of neural network is also known as a soft committee machine in the literature [92].

6 Concluding Perspectives

Non-equilibrium statistical physics has made great strides over the last twenty years. Stochastic thermodynamics now provides a mature framework to study small systems far from equilibrium based on the mathematical theory of Markov processes. We can use it to analyse the thermodynamic properties of systems that process information and where fluctuations play an important role, arbitrarily far from equilibrium. This framework serves as the theoretical basis for a systematic study of the three steps of biological information processing: sensing, building a model, and generalising from examples.

6.1 Thermodynamic bounds on learning

In this thesis, we have used concepts and tools from stochastic thermodynamics, statistical physics and information theory to study learning in neural networks with thermal fluctuations from a thermodynamic perspective. Focusing on the last two steps of information processing, we obtained three key results. First, we saw that the amount of information a neural network can extract from data to build a model for that data is bounded by the thermodynamic costs of learning. Second, we showed that the total entropy production of a network also constrains its ability to infer a rule from examples. This further implies a trade-off between the ability of such a network to generalise from examples to previously unseen inputs and its total dissipation during learning.

However, learning a rule once is not enough in the ever-changing environment that biological systems face. It is therefore plausible that biological systems are also optimised with respect to the steady state that ensues when a learned model is continuously updated to reflect patterns emerging in newly acquired data. We have therefore studied the steady state behaviour of a neural network and found a general trade-off where an undesirable quantity, the total entropy production of the network, bounds the product of two desired quantities, namely the speed and the reliability of learning. This is the third key result of this thesis.

Our results are based on the second law of thermodynamics and hold for both supervised and unsupervised learning problems. They apply to batch and online learning and we have illustrated them for a number of different learning algorithms. They hold at all times and do not require taking the thermodynamic limit.

6.2 Classical computation and beyond

Our work opens up a number of avenues for further research. An intriguing challenge is to design learning algorithms which optimise thermodynamic efficiency. Learning rules which optimise information extraction during online learning have been found previously [147, 148]. Finding an algorithm that optimises the efficiency looks like a more serious challenge not least because the (excess) heat that enters the efficiency is a function of the entire trajectory and not just the state, making the problem non-local in time.

Another important extension of our work concerns more complicated neural architectures. While our fluctuation theorem and the general bound of Chapter 5 impose thermodynamic constraints on the ability of multi-layer neural networks to learn, it is tempting to speculate which types of architecture are the most efficient. The capabilities of networks with just a single layer are already remarkable: a network of binary neurons ($\sigma = \pm 1$) with just a single intermediate layer can implement *any* Boolean function of its N inputs and requires at most 2^N intermediate units [149], while a network with continuous neurons ($\sigma \in \mathbb{R}$, *e.g.* $\sigma = \tanh(\mathcal{A})$) is capable of approximating any continuous function of its inputs to any required accuracy if the number of intermediate units is not constrained [150, 151]. However, the number of hidden units grows very large in these networks, while multi-layer networks can achieve similar results with far fewer hidden units in each layer [124, 152]. Biological networks, such as the human retina [3], seem to prefer several layers of neurons with discrete receptive fields, where each neuron in the intermediate layers is only wired to a subset of neurons in the previous layer, rather than being connected to all of the inputs of the network. Elucidating the extent to which thermodynamic efficiency played a role in the evolution of these architectures is an intriguing challenge for the future.

While we have focused on biologically motivated and thus classical models of neural networks, the prospect of quantum computing [153] promises to upend many of the existing paradigms of computing, with machine learning being no exception [154]. The thermodynamics of quantum few-particle systems are currently also drawing a lot of interest [155]. Having this in mind, we note that the results of stochastic thermodynamics are directly applicable to driven or open quantum systems whenever coherences can be ignored [20] (but see Ref. [156] for a review focused on open quantum systems). A fascinating question is hence whether coherences between the weights can improve not just the computational prowess of quantum neural networks, but also their thermodynamic efficiency. In this context, it is important to keep in mind that coherences universally *reduce* the efficiency of quantum heat engines in the linear response regime [157]. Analysing the trade-off between the additional computational capabilities that coherence offers and the dissipative losses it inevitably incurs is an exciting topic for further research.

6.3 Constraining computation: energy, time and data

We have considered the limitations on computation in neural networks that are a consequence of the second law of thermodynamics and account for the free energy costs of computations. The resulting class of bounds is intriguing from a conceptual point of view and particularly important for small, fluctuating systems far from equilibrium.

However, there are further limiting factors on the ability to compute and to perform inference. One is the availability of data: for a given inference problem, such as inferring a teacher $\boldsymbol{T}$ from a number of samples $(\sigma_{\mathrm{T}}, \boldsymbol{\xi})$, there is a minimum amount of data that is required to make any prediction that is better than simply flipping a coin. A second limiting factor is time: ideally, we would like to have an algorithm whose completion time scales polynomially with the system size, rather than exponentially. It has recently become clear that these two constraints can be understood using statistical mechanics in terms of phase transitions [158]. It will be interesting to see whether the thermodynamic limits that we have derived in this thesis fit into this picture, and if so, where they can be found.

Bibliography

1. Bialek, W. *Biophysics : Searching for Principles* (Princeton University Press, 2011).
2. Alberts, B. *et al. Molecular Biology of the Cell* (Garland Science, New York, 2008).
3. Kandel, E. R., Schwartz, J. H., Jessell, T. M. *et al. Principles of Neural Science* (McGraw-Hill New York, 2000).
4. Barkai, N. & Leibler, S. Robustness in simple biochemical networks. *Nature* **387,** 913–917 (1997).
5. Hopfield, J. J. Kinetic Proofreading: A New Mechanism for Reducing Errors in Biosynthetic Processes Requiring High Specificity. *Proc. Natl. Acad. Sci.* **71,** 4135–4139 (1974).
6. Ninio, J. Kinetic amplification of enzyme discrimination. *Biochimie* **57,** 587–595 (1975).
7. Bennett, C. H. Dissipation-error tradeoff in proofreading. *Biosystems* **11,** 85–91 (1979).
8. MacKay, D. J. *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, 2003).
9. Lan, G., Sartori, P., Neumann, S., Sourjik, V. & Tu, Y. The energy–speed–accuracy trade-off in sensory adaptation. *Nat. Phys.* **8,** 422–428 (2012).
10. Sterling, P. & Laughlin, S. *Principles of Neural Design* (MIT Press, 2015).
11. Lahiri, S., Sohl-Dickstein, J. & Ganguli, S. A universal tradeoff between power, precision and speed in physical communication. arXiv: `1603.07758` (2016).
12. Chandler, D. *Introduction to Modern Statistical Mechanics* (Oxford University Press, 1987).
13. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* **381,** 520–522 (1996).
14. Kubo, R., Toda, M. & Hashitsume, N. *Statistical Physics II - Nonequilibrium* (Springer, 1985).

15. Pottier, N. *Nonequilibrium Statistical Physics: Linear Irreversible Processes* (Oxford University Press, Oxford, 2010).
16. Boltzmann, L. Über die Beziehung zwischen dem zweiten Hauptsatze der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung resp. den Sätzen über das Wärmegleichgewicht. *Wiener Berichte* **76,** 373–435 (1877).
17. Maxwell, J. C. *Theory of Heat* (Appleton, London, 1871).
18. Gibbs, J. W. *Elementary Principles in Statistical Mechanics* 207 (Charles Scribner's Sons, 1902).
19. Bressloff, P. *Stochastic Processes in Cell Biology* (Springer, 2014).
20. Seifert, U. Stochastic thermodynamics, fluctuation theorems and molecular machines. *Rep. Prog. Phys.* **75,** 126001 (2012).
21. Schmiedl, T. & Seifert, U. Efficiency at maximum power: An analytically solvable model for stochastic heat engines. *Europhys. Lett.* **81,** 20003 (2008).
22. Blickle, V. & Bechinger, C. Realization of a micrometre-sized stochastic heat engine. *Nat. Phys.* **8,** 143–146 (2011).
23. Callen, H. *Thermodynamics and an Introduction to Thermostatistics* 2nd (John Wiley & Sons, New York, 1985).
24. Abah, O. *et al.* Single-Ion Heat Engine at Maximum Power. *Phys. Rev. Lett.* **109,** 203006 (2012).
25. Rossnagel, J. *et al.* A single-atom heat engine. *Science* **352,** 325–329 (2016).
26. Toyabe, S. *et al.* Nonequilibrium Energetics of a Single F1-ATPase Molecule. *Phys. Rev. Lett.* **104,** 198103 (2010).
27. Toyabe, S., Watanabe-Nakayama, T., Okamoto, T., Kudo, S. & Muneyuki, E. Thermodynamic efficiency and mechanochemical coupling of F1-ATPase. *Proc. Natl. Acad. Sci.* **108,** 17951–17956 (2011).
28. Zimmermann, E. & Seifert, U. Efficiencies of a molecular motor: a generic hybrid model applied to the F 1 -ATPase. *New J. Phys.* **14,** 103023 (2012).
29. Sarikaya, M., Tamerler, C., Jen, A., Schulten, K. & Baneyx, F. Molecular biomimetics: nanotechnology through biology. *Nat. Mater.* **2,** 577–585 (2003).
30. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).
31. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (John Wiley & Sons, 2006).
32. Landauer, R. Information is Physical. *Phys. Today* **44,** 23–29 (1991).

33. *Maxwell's Demon: Entropy, Information, Computing* 2nd ed. (eds Leff, H. S. & Rex, A. F.) (IOP Publishing, 2003).

34. Maruyama, K., Nori, F. & Vedral, V. Colloquium: The physics of Maxwell's demon and information. *Rev. Mod. Phys.* **81,** 1–23 (2009).

35. Jacobs, K. Second law of thermodynamics and quantum feedback control: Maxwell's demon with weak measurements. *Phys. Rev. A* **80,** 012322 (2009).

36. Sagawa, T. & Ueda, M. Generalized Jarzynski Equality under Nonequilibrium Feedback Control. *Phys. Rev. Lett.* **104,** 090602 (2010).

37. Mandal, D. & Jarzynski, C. Work and information processing in a solvable model of Maxwell's demon. *Proc. Natl. Acad. Sci.* **109,** 11641–11645 (2012).

38. Hartich, D., Barato, A. C. & Seifert, U. Stochastic thermodynamics of bipartite systems: transfer entropy inequalities and a Maxwell's demon interpretation. *J. Stat. Mech.* **2014,** P02016 (2014).

39. Horowitz, J. M. & Esposito, M. Thermodynamics with Continuous Information Flow. *Phys. Rev. X* **4,** 031015 (2014).

40. Parrondo, J. M. R., Horowitz, J. M. & Sagawa, T. Thermodynamics of information. *Nat. Phys.* **11,** 131–139 (2015).

41. Toyabe, S., Sagawa, T., Ueda, M., Muneyuki, E. & Sano, M. Experimental demonstration of information-to-energy conversion and validation of the generalized Jarzynski equality. *Nat. Phys.* **6,** 988–992 (2010).

42. Landauer, R. Irreversibility and Heat Generation in the Computing Process. *IBM J. Res. Dev.* **5,** 183–191 (1961).

43. Bérut, A. *et al.* Experimental verification of Landauer's principle linking information and thermodynamics. *Nature* **483,** 187–189 (2012).

44. Jun, Y., Gavrilov, M. & Bechhoefer, J. High-Precision Test of Landauer's Principle in a Feedback Trap. *Phys. Rev. Lett.* **113,** 190601 (2014).

45. Berg, H. C. & Purcell, E. M. Physics of chemoreception. *Biophys. J.* **20,** 193–219 (1977).

46. Qian, H. & Reluga, T. C. Nonequilibrium Thermodynamics and Nonlinear Kinetics in a Cellular Signaling Switch. *Phys. Rev. Lett.* **94,** 028101 (2005).

47. Keymer, J. E., Endres, R. G., Skoge, M., Meir, Y. & Wingreen, N. S. Chemosensing in Escherichia coli: Two regimes of two-state receptors. *Proc. Natl. Acad. Sci.* **103,** 1786–1791 (2006).

48. Tu, Y. The nonequilibrium mechanism for ultrasensitivity in a biological switch: Sensing by Maxwell's demons. *Proc. Natl. Acad. Sci.* **105,** 11737–11741 (2008).

49. Endres, R. G. & Wingreen, N. S. Maximum likelihood and the single receptor. *Phys. Rev. Lett.* **103,** 158101 (2009).
50. Mehta, P. & Schwab, D. J. Energetic costs of cellular computation. *Proc. Natl. Acad. Sci.* **109,** 17978–17982 (2012).
51. Skoge, M., Naqvi, S., Meir, Y. & Wingreen, N. S. Chemical Sensing by Nonequilibrium Cooperative Receptors. *Phys. Rev. Lett.* **110,** 248102 (2013).
52. Govern, C. C. & ten Wolde, P. R. Energy Dissipation and Noise Correlations in Biochemical Sensing. *Phys. Rev. Lett.* **113,** 258102 (2014).
53. Govern, C. C. & ten Wolde, P. R. Optimal resource allocation in cellular sensing systems. *Proc. Natl. Acad. Sci.* **111,** 17486–17491 (2014).
54. Barato, A. C., Hartich, D. & Seifert, U. Efficiency of cellular information processing. *New J. Phys.* **16,** 103024 (2014).
55. Lang, A. H., Fisher, C. K., Mora, T. & Mehta, P. Thermodynamics of Statistical Inference by Cells. *Phys. Rev. Lett.* **113,** 148103 (2014).
56. Sartori, P., Granger, L., Lee, C. F. & Horowitz, J. M. Thermodynamic Costs of Information Processing in Sensory Adaptation. *PLoS Comput. Biol.* **10,** e1003974 (2014).
57. Ito, S. & Sagawa, T. Maxwell's demon in biochemical signal transduction with feedback loop. *Nat. Commun.* **6,** 7498 (2015).
58. Ten Wolde, P. R., Becker, N. B., Ouldridge, T. E. & Mugler, A. Fundamental Limits to Cellular Sensing. *J. Stat. Phys.* **162,** 1395–1424 (2016).
59. Andrieux, D. & Gaspard, P. Nonequilibrium generation of information in copolymerization processes. *Proc. Natl. Acad. Sci.* **105,** 9516–9521 (2008).
60. Murugan, A., Huse, D. A. & Leibler, S. Speed, dissipation, and error in kinetic proofreading. *Proc. Natl. Acad. Sci.* **109,** 12034–12039 (2012).
61. Hartich, D., Barato, A. C. & Seifert, U. Nonequilibrium sensing and its analogy to kinetic proofreading. *New J. Phys.* **17,** 055026 (2015).
62. Lahiri, S., Wang, Y., Esposito, M. & Lacoste, D. Kinetics and thermodynamics of reversible polymerization in closed systems. *New J. Phys.* **17,** 085008 (2015).
63. Barato, A. C. & Seifert, U. Thermodynamic Uncertainty Relation for Biomolecular Processes. *Phys. Rev. Lett.* **114,** 158101 (2015).
64. Sekimoto, K. Kinetic Characterization of Heat Bath and the Energetics of Thermal Ratchet Models. *J. Phys. Soc. Japan* **66,** 1234–1237 (1997).
65. Sekimoto, K. Langevin Equation and Thermodynamics. *Prog. Theor. Phys. Suppl.* **130,** 17–27 (1998).

66. Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **78,** 2690–2693 (1997).

67. Jarzynski, C. Equilibrium free-energy differences from nonequilibrium measurements: A master-equation approach. *Phys. Rev. E* **56,** 5018–5035 (1997).

68. Evans, D. J., Cohen, E. G. D. & Morriss, G. P. Probability of second law violations in shearing steady states. *Phys. Rev. Lett.* **71,** 2401–2404 (1993).

69. Gallavotti, G. & Cohen, E. G. D. Dynamical Ensembles in Nonequilibrium Statistical Mechanics. *Phys. Rev. Lett.* **74,** 2694–2697 (1995).

70. Kurchan, J. Fluctuation theorem for stochastic dynamics. *J. Phys. A: Math. Gen.* **31,** 3719–3729 (1998).

71. Lebowitz, J. L. & Spohn, H. A Gallavotti–Cohen-Type Symmetry in the Large Deviation Functional for Stochastic Dynamics. *J. Stat. Phys.* **95,** 333–365 (1999).

72. Seifert, U. Stochastic thermodynamics: Principles and perspectives. *Eur. Phys. J. B* **64,** 423–431 (2008).

73. Van Kampen, N. G. *Stochastic Processes in Physics and Chemistry* (Elsevier, 1992).

74. Gardiner, C. *Stochastic Methods* (Springer, Berlin-Heidelberg-New York-Tokyo, 2009).

75. Risken, H. *The Fokker-Planck Equation* (Springer, 1996).

76. Bustamante, C., Liphardt, J. & Ritort, F. The nonequilibrium thermodynamics of small systems. *Phys. Today* **58,** 43–48 (2005).

77. Jarzynski, C. Equalities and Inequalities: Irreversibility and the Second Law of Thermodynamics at the Nanoscale. *Annu. Rev. Condens. Matter Phys.* **2,** 329–351 (2011).

78. Sekimoto, K. *Stochastic Energetics* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2010).

79. Van Den Broeck, C. & Esposito, M. Ensemble and trajectory thermodynamics: A brief introduction. *Physica A* **418,** 6–16 (2015).

80. Seifert, U. Stochastic thermodynamics: From principles to the cost of precision. *Phys. A Stat. Mech. its Appl.* 1–10 (2017).

81. Hartich, D. *Stochastic thermodynamics of information processing: bipartite systems with feedback, signal inference and information storage* PhD thesis (Universität Stuttgart, 2017).

82. Seifert, U. Entropy Production along a Stochastic Trajectory and an Integral Fluctuation Theorem. *Phys. Rev. Lett.* **95,** 040602 (2005).

83. Tribus, M. & McIrvine, E. C. Energy and information. *Sci. Am.* **225,** 179–190 (1971).

84. Horowitz, J. M. Multipartite information flow for multiple Maxwell demons. *J. Stat. Mech.* **2015,** P03006 (2015).

85. Allahverdyan, A. E., Janzing, D. & Mahler, G. Thermodynamic efficiency of information and heat flow. *J. Stat. Mech.* **2009,** P09011 (2009).

86. Oono, Y. & Paniconi, M. Steady State Thermodynamics. *Prog. Theor. Phys. Suppl.* **130,** 29–44 (1998).

87. Hatano, T. & Sasa, S.-i. Steady-State Thermodynamics of Langevin Systems. *Phys. Rev. Lett.* **86,** 3463–3466 (2001).

88. Speck, T. & Seifert, U. Integral fluctuation theorem for the housekeeping heat. *J. Phys. A: Math. Gen.* **38,** L581–L588 (2005).

89. Trepagnier, E. H. *et al.* Experimental test of Hatano and Sasa's nonequilibrium steady-state equality. *Proc. Natl. Acad. Sci.* **101,** 15038–15041 (2004).

90. Van den Broeck, C. & Esposito, M. Three faces of the second law. II. Fokker-Planck formulation. *Phys. Rev. E* **82,** 011144 (2010).

91. Goldt, S. & Seifert, U. Stochastic Thermodynamics of Learning. *Phys. Rev. Lett.* **118,** 010601 (2017).

92. Engel, A. & Van den Broeck, C. *Statistical Mechanics of Learning* (Cambridge University Press, 2001).

93. Newsome, W. T., Britten, K. H. & Movshon, J. A. Neuronal correlates of a perceptual decision. *Nature* **341,** 52–4 (1989).

94. Marr, D. A theory of cerebellar cortex. *J. Physiol.* **202,** 437–470 (1969).

95. Albus, J. S. A theory of cerebellar function. *Math. Biosci.* **10,** 25–61 (1971).

96. Dayan, P. & Abbott, L. F. *Theoretical Neuroscience* (MIT Press, 2001).

97. Hebb, D. O. *The organization of behavior: A neuropsychological approach* (John Wiley & Sons, 1949).

98. Abreu, D. & Seifert, U. Extracting work from a single heat bath through feedback. *Europhys. Lett.* **94,** 10001 (2011).

99. Gardner, E. Maximum Storage Capacity in Neural Networks. *Europhys. Lett.* **4,** 481–485 (1987).

100. Hartich, D., Barato, A. C. & Seifert, U. Sensory capacity: An information theoretical measure of the performance of a sensor. *Phys. Rev. E* **93,** 022116 (2016).

101. Brunel, N., Hakim, V., Isope, P., Nadal, J.-P. & Barbour, B. Optimal Information Storage and the Distribution of Synaptic Weights. *Neuron* **43,** 745–757 (2004).

102. Barbour, B., Brunel, N., Hakim, V. & Nadal, J.-P. What can we learn from synaptic weight distributions? *Trends Neurosci.* **30,** 622–629 (2007).

103. Tkačik, G., Walczak, A. M. & Bialek, W. Optimizing information flow in small genetic networks. *Phys. Rev. E* **80,** 031920 (2009).

104. Sokolowski, T. R. & Tkačik, G. Optimizing information flow in small genetic networks. IV. Spatial coupling. *Phys. Rev. E* **91,** 062710 (2015).

105. Horowitz, J. M. & Sandberg, H. Second-law-like inequalities with information and their interpretations. *New J. Phys.* **16,** 125007 (2014).

106. Vallet, F. The Hebb Rule for Learning Linearly Separable Boolean Functions: Learning and Generalization. *Europhys. Lett.* **8,** 747–751 (1989).

107. Rosenblatt, F. *Principles of Neurodynamics* (Spartan, New York, 1962).

108. Anlauf, J. K. & Biehl, M. The AdaTron: An Adaptive Perceptron Algorithm. *Europhys. Lett.* **10,** 687–692 (1989).

109. Goldt, S. & Seifert, U. Thermodynamic efficiency of learning a rule in neural networks. *New J. Phys.* **19,** 113001 (2017).

110. Ito, M. & Kano, M. Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex. *Neurosci. Lett.* **33,** 253–258 (1982).

111. Ito, M., Sakurai, M. & Tongroach, P. Climbing fibre induced depression of both mossy fibre responsiveness and glutamate sensitivity of cerebellar Purkinje cells. *J. Physiol.* **324,** 113–134 (1982).

112. Biehl, M. & Riegler, P. On-Line Learning with a Perceptron. *Europhys. Lett.* **28,** 525–530 (1994).

113. Mace, C. W. H. & Coolen, A. C. C. Statistical mechanical analysis of the dynamics of learning in perceptrons. *Stat. Comput.* **8,** 55–88 (1998).

114. Heskes, T. M. & Kappen, B. Learning processes in neural networks. *Phys. Rev. A* **44,** 2718–2726 (1991).

115. *NIST Handbook of Mathematical Functions* (eds Olver, F. W. J., Lozier, D. W., Boisvert, R. F. & Clark, C. W.) (Cambridge University Press, New York, NY, 2010).

116. Minsky, M. L. & Papert, S. A. *Perceptrons* (MIT Press, Cambridge MA, Cambridge, 1969).

117. Barkai, N., Seung, H. S. & Sompolinsky, H. Local and Global Convergence of On-Line Learning. *Phys. Rev. Lett.* **75,** 1415–1418 (1995).

118. Polya, G. *Induction and Analogy in Mathematics* (Princeton University Press, 1954).

119. Cover, T. M. Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition. *IEEE Trans. Electron. Comput.* **EC-14,** 326–334 (1965).

120. Kinouchi, O. & Caticha, N. Lower bounds on generalization errors for drifting rules. *J. Phys. A: Math. Gen.* **26,** 6161–6171 (1993).

121. Biehl, M. & Schwarze, H. Learning drifting concepts with neural networks. *J. Phys. A: Math. Gen.* **26,** 2651–2665 (1993).

122. Colabrese, S., Gustavsson, K., Celani, A. & Biferale, L. Flow Navigation by Smart Microswimmers via Reinforcement Learning. *Phys. Rev. Lett.* **118,** 158004 (2017).

123. Jaynes, E. T. Information Theory and Statistical Mechanics. *Phys. Rev.* **106,** 620–630 (1957).

124. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).

125. Jensen, F. V. & Nielsen, T. D. *Bayesian Networks and Decision Graphs* Second (Springer, 2007).

126. Van den Broeck, C. & Reimann, P. Unsupervised Learning by Examples: On-line versus Off-line. *Phys. Rev. Lett.* **76,** 2188–2191 (1996).

127. Hinton, G. E., Osindero, S. & Teh, Y. W. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* **18,** 1527–1554 (2006).

128. Chernyak, V. Y., Chertkov, M. & Jarzynski, C. Path-integral analysis of fluctuation theorems for general Langevin processes. *J. Stat. Mech.* **2006,** P08001–P08001 (2006).

129. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323,** 533–536 (1986).

130. Smolensky, P. *Information processing in dynamical systems: Foundations of harmony theory* (MIT Press, 1986).

131. Hinton, G. E. Training Products of Experts by Minimizing Contrastive Divergence. *Neural Comput.* **14,** 1771–1800 (2002).

132. Tieleman, T. Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. *Proc. 25th Int. Conf. Mach. Learn.* **307,** 7 (2008).

133. Kuh, A., Petsche, T. & Rivest, R. L. Learning Time-varying Concepts. *Adv. Neural Inf. Process. Syst.* 183–189 (1991).

134. Ito, S. & Sagawa, T. Information Thermodynamics on Causal Networks. *Phys. Rev. Lett.* **111,** 180603 (2013).

135. Schreiber, T. Measuring Information Transfer. *Phys. Rev. Lett.* **85,** 461–464 (2000).

136. Opper, M. On-line versus Off-line Learning from Random Examples: General Results. *Phys. Rev. Lett.* **77,** 4671–4674 (1996).

137. Esposito, M. & Van Den Broeck, C. Three faces of the second law. I. Master equation formulation. *Phys. Rev. E* **82,** 011143 (2010).

138. Pietzonka, P., Barato, A. C. & Seifert, U. Universal bounds on current fluctuations. *Phys. Rev. E* **93,** 052145 (2016).

139. Gingrich, T. R., Horowitz, J. M., Perunov, N. & England, J. L. Dissipation Bounds All Steady-State Current Fluctuations. *Phys. Rev. Lett.* **116,** 120601 (2016).

140. Pietzonka, P., Ritort, F. & Seifert, U. Finite-time generalization of the thermodynamic uncertainty relation. *Phys. Rev. E* **96,** 012101 (2017).

141. Horowitz, J. M. & Gingrich, T. R. Proof of the finite-time thermodynamic uncertainty relation for steady-state currents. *Phys. Rev. E* **96,** 020103 (2017).

142. Gingrich, T. R., Rotskoff, G. M. & Horowitz, J. M. Inferring dissipation from current fluctuations. *J. Phys. A: Math. Theor.* **50,** 184004 (2017).

143. Shiraishi, N. Finite-time thermodynamic uncertainty relation do not hold for discrete-time Markov process. arXiv: `1706.00892` (2017).

144. Proesmans, K. & Van den Broeck, C. Discrete-time thermodynamic uncertainty relation. *Europhys. Lett.* **119,** 20001 (2017).

145. Polettini, M., Lazarescu, A. & Esposito, M. Tightening the uncertainty principle for stochastic currents. *Phys. Rev. E* **94,** 1–10 (2016).

146. Maes, C. Frenetic Bounds on the Entropy Production. *Phys. Rev. Lett.* **119,** 160601 (2017).

147. Kinouchi, O. & Caticha, N. Biased learning in Boolean perceptrons. *Physica A* **185,** 411–416 (1992).

148. Kinouchi, O. & Caticha, N. On-line versus off-line learning in the linear perceptron: A comparative study. *Phys. Rev. E* **52,** 2878–2886 (1995).

149. Denker, J. *et al.* Large Automatic Learning, Rule Extraction, and Generalization. *Complex Syst.* **1,** 877–922 (1987).

150. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Math. Control. Signals, Syst.* **2,** 303–314 (1989).

151. Hornik, K., Stinchcombe, M. & White, H. Multilayer feedforward networks are universal approximators. *Neural Networks* **2,** 359–366 (1989).

152. Bengio, Y. Learning Deep Architectures for AI. *Found. Trends® Mach. Learn.* **2,** 1–127 (2009).

153. Nielsen, M. A. & Chuang, I. L. *Quantum Computation and Quantum Information* (Cambridge University Press, Cambridge, 2010).

154. Biamonte, J. *et al.* Quantum machine learning. *Nature* **549,** 195–202 (2017).

155. Vinjanampathy, S. & Anders, J. Quantum thermodynamics. *Contemp. Phys.* **57,** 545–579 (2016).

156. Esposito, M., Harbola, U. & Mukamel, S. Nonequilibrium fluctuations, fluctuation theorems, and counting statistics in quantum systems. *Rev. Mod. Phys.* **81,** 1665–1702 (2009).

157. Brandner, K., Bauer, M. & Seifert, U. Universal Coherence-Induced Power Losses of Quantum Heat Engines in Linear Response. *Phys. Rev. Lett.* **119,** 170602 (2017).

158. Zdeborová, L. & Krzakala, F. Statistical physics of inference: thresholds and algorithms. *Adv. Phys.* **65,** 453–552 (2016).

Danksagung

Zum Abschluss dieser Arbeit möchte ich mich bei den Personen bedanken, deren Unterstützung ich während der Promotion erfahren durfte.

Mein erster Dank gilt natürlich Herrn Professor Seifert, der mir an seinem Institut die Gelegenheit zu dieser Promotion gab und mir nicht nur großen Freiraum gewährte, sondern stets auch mit wertvollen Ratschlägen zur Seite stand.

Weiterhin möchte ich Herrn Professor Holm für die Anfertigung des Zweitberichtes und Frau Professor Barz für die Übernahme des Prüfungsvorsitzes danken.

Frau Steinhauser war mir von Anfang an eine unverzichtbare Hilfe, sei es im Gespräch mit verschiedenen Behörden oder im Alltag am Institut. Dafür möchte ich ihr herzlich danken.

Ein besonderer Dank gilt Dr. David Hartich für die unzähligen Gespräche, die mir immer eine große Hilfe und stets ein besonderes Vergnügen waren. Allen meinen Kollegen am Institut danke ich für die angenehme und stets kollegiale Zusammenarbeit und für den regen Austausch, insbesondere mit Patrick Pietzonka und Dr. Kay Brandner.

Den Administratoren Dr. Timo Bihr, Robert Wulfert, Dr. Michael Bauer, Matthias Uhl und Lukas Fischer danke ich für die schnelle und unkomplizierte Hilfe bei allen IT-Problemen.

Mein größter Dank gilt schließlich meinen Eltern, die mich über viele Jahre hinweg in allen Lebenslagen immer unterstützt, ermutigt und inspiriert haben und mir dabei stets alle Freiheiten ließen.

Ehrenwörtliche Erklärung

Ich erkläre, dass ich diese Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Stuttgart, den 19. Dezember 2017 Sebastian Goldt