Appendix A

Information in Deep Learning

> *What can we demand from any physical theory? . . . nature [is] a difficult problem, but not a mystery for the human mind.*
>
> Ludwig Boltzmann [76]

In our *Initialization*, §0, we introduced our effective theory approach to understanding neural networks via the lens of theoretical physics. In particular, we discussed how thermodynamics was used to clarify the behavior of artificial machines such as the steam engine, and then we described how statistical mechanics was developed to explain how these macroscopic laws arise from the statistical behavior of many microscopic elements. With this perspective, we suggested that a similar framework might be applied to the difficult problem of deep learning theory, which we have now demystified from *Pretraining* all the way to the *End of Training*, §1 to §∞.

In this first appendix, we'll make the connection between deep learning and these fields of physics even more detailed. To do so, we'll reformulate a few of our previous results in terms of **information theory**. Initially formalized by Shannon as a way of quantitatively understanding digital communication, information theory was developed about half a century after statistical mechanics and is the statistical microscopic theory most fitting for the digital Information Age.

Although they a priori consider very different subject matter, both statistical mechanics and information theory share a joint language and ultimate focus on the same main fundamental concept: *entropy*. A particularly nice organization of entropies defines the *mutual information*, a positive quantity that can be used to characterize how much the measurement of one observable can inform us about another. This gives a nice way to quantify the overall statistical dependence due to the interactions of random variables. As such, the first section of this appendix (§A.1) will give a self-contained introduction to entropy and mutual information, making sure to point out the connections between the very physical and analog setting of statistical mechanics and the very abstract and digital world of information theory.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

With these new tools, we will be able to further understand information in deep learning. In particular, for infinite-width neural networks (§A.2), we'll find new perspectives on the noninteraction of neurons within a layer and on the necessity of criticality for preserving the information essential for distinguishing between input samples in deep networks. Then, at finite width (§A.3) we'll use a *variational principle* to demonstrate how the *principle of maximum entropy* for nearly-Gaussian distributions enables us to compute entropies and informations up to order $1/n^3$ while only needing to know the effective ℓ-th-layer preactivation distribution truncated at order $1/n$.

At order $1/n^2$, this will let us see nonzero mutual information between groups of neurons in a layer that grows quadratically with depth, further quantifying the interaction of neurons at finite width and providing an information-theoretic interpretation and generalization of our Hebbian learning result from §6.4.1.

At order $1/n^3$, we will see how to pick a depth-to-width ratio, $r \equiv L/n$, that maximizes the mutual information as a functional of the activation function. This optimal aspect ratio, $r^\star$, arises from an unsupervised learning objective and defines an activation-function-dependent scale that separates *effectively-deep* networks – that perform well – from *overly-deep* networks – that are no longer trainable.

Also at order $1/n^3$, a generalization of the mutual information for three groups of neurons will show that the information in a finite-width layer is stored redundantly between neurons. This analysis provides a new perspective on the coarse-graining mechanism of *RG flow* by enabling us to understand how the information from inputs gets represented by, and shared among, the deeper-layer neurons.

Finally, note that while most of the computations presented here focus on the prior distribution of preactivations as a means of investigating the inductive bias of the network architecture and activation function, these information-theoretic tools naturally extend to the various joint distributions we've considered throughout the book as well as the Bayesian posterior distribution and the complicated distributions of fully-trained networks. This leaves many more things to be computed, and so we hope that this chapter provides a useful introduction to a new toolkit that can be used for furthering your understanding of deep learning. In our second and final appendix, §B, we'll give a *residual* example to demonstrate this in the setting of *residual networks*.

A.1 Entropy and Mutual Information

In this section we give a very brief overview of the concepts of entropy and mutual information that play essential roles both in *statistical mechanics* [77] and *information theory* [78, 79].

Let's start with a discrete random variable $x \in \mathcal{X}$ governed by the probability distribution $p(x)$. For this discussion, it is nice to think of x as a particular observable outcome and $\mathcal{X}$ as the set of possible outcomes. The **entropy** of a probability distribution is given by

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

$$\mathcal{S}[p(x)] \equiv -\sum_{x\in\mathcal{X}} p(x)\log p(x), \tag{A.1}$$

which is a *functional* of the distribution, taking a distribution as an argument and outputting a number. Thus, we should think of the entropy as a property of an entire probability distribution.

To gain some intuition for why this quantity could be useful, let's consider its two extremes. First, when the distribution is perfectly *ordered* – that is, $p(x) = \delta_{xx'}$ such that $p(x') = 1$ with *absolute certainty* for a particular outcome, $x' \in \mathcal{X}$, and zero for all others – then the entropy is minimal, given by

$$\mathcal{S}[p(x)] = -\sum_{x\in\mathcal{X}} \delta_{xx'}\log(\delta_{xx'}) = 0, \tag{A.2}$$

since x' contributes $1\log(1) = 0$, and the other values of x contribute $0\log(0) = 0$. Second, when the distribution is completely *disordered* – that is, $p(x) = 1/|\mathcal{X}|$, such that the possible outcomes x are distributed uniformly and no outcome is more likely than any other – then entropy is maximal, given by

$$\mathcal{S}[p(x)] = -\sum_{x\in\mathcal{X}} \frac{1}{|\mathcal{X}|}\log\left(\frac{1}{|\mathcal{X}|}\right) = \log(|\mathcal{X}|). \tag{A.3}$$

For this reason, in physics the entropy is often regarded as a measure of the **disorder** of a system characterized by a random variable x.[1] In this maximal case when each outcome is equally likely, the entropy can also be interpreted as a way of counting the

[1]Since we're currently wearing our physicist hats, one remark is in (dis)order for the units of the entropy. As per our discussion of *dimensional analysis* in footnote 15 of §1.3, the logarithm of a probability has the same units as the *action*, and for the same reason must be dimensionless. Nevertheless, by changing the base of the logarithm, we can change the multiplicative constant in front of (A.1) and thus change the meaning of the entropy:

- With our physicist hats still on, we would use the natural logarithm with base e, which measures entropy in units with a very silly name called *nats*. (Some physicists also multiply the expression (A.1) by the Boltzmann constant k_{B}, as is natural in macroscopic applications of the entropy in *thermodynamics*, which gives the entropy the unit of *joules per kelvin*.)
- With our computer scientist hats on, which we put on in anticipation of the next paragraph in the main text, we would use the logarithm with base 2, which measures units in *binary digits* or the hopefully familiar **bits**. This is most natural in an information theory context, for which the entropy (A.1) is sometimes called the *Shannon entropy*. The reason bits are also used as the units for computer memory is due to the counting-of-states intuition for the entropy: a hard drive that can store 10^9 `bits` has 2^{10^9} unique states, with a priori equal plausibility for any particular arrangement of those bits, i.e., for any particular state of the system.

In this chapter, we'll use the natural base e, which is also very natural when studying the entropy of Gaussian distributions and nearly-Gaussian distributions.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

number of states of a system – i.e., the number of distinct outcomes in $\mathcal{X}$ – since it's equal to the logarithm of such a count.[2]

To gain even more intuition, let's consider a perspective from information theory. In the entropy formula (A.1), the quantity inside the expectation is called the **surprisal**:

$$s(x) \equiv -\log p(x) = \log\left[\frac{1}{p(x)}\right] ; \tag{A.4}$$

since for a particular outcome, x, the distribution $p(x)$ quantifies the frequency or plausibility of observing x, and since the logarithm is monotonic in its argument, the surprisal grows with the a priori rareness of the outcome x and hence quantifies how surprising or informative actually observing a particular x is. As such, the surprisal is also a quantitative measure of how much new **information** is gained after making such an observation. Averaging the surprisal over all possible outcomes, $\mathcal{X}$, gives the entropy (A.1), which can thus be understood as the expected surprise or amount of information to be gained from making an observation:

$$\mathcal{S}[p(x)] \equiv \mathbb{E}\left[s(x)\right]. \tag{A.5}$$

In addition to admitting various nice interpretations, the entropy also obeys a few nice mathematical properties. To start, it is manifestly nonnegative:

$$\mathcal{S}[p(x)] \geq 0. \tag{A.6}$$

You can see this by noting that the allowed range of a probability, $0 \leq p(x) \leq 1$, implies the nonnegativity of the quantity $-p(x)\log p(x) \geq 0$, which in turn implies that the entropy (A.1) is a sum of nonnegative terms. Moreover, the entropy takes its minimum value and vanishes, (A.2), if and only if the distribution is perfectly ordered, given by a Kronecker delta for one particular outcome.

[2] This connection between the uniform distribution over a finite number of states and the maximum of the entropy is an example of the *principle of maximum entropy* and descends from what's called Laplace's *principle of indifference*: without any other information, all the a priori probabilities for a system to be in any particular state should be equal. (As we will discuss in §A.3, when we do have some information about the state, then the entropy is no longer maximized by a uniform distribution.)

Note that this indifference principle is applied to macroscopic thermodynamic states of (nearly-)equal energy as a principal assumption of microscopic statistical mechanics, and the associated entropy, (A.3), is sometimes called the *Boltzmann entropy*. In contrast, the fully-general formula (A.1) is sometimes called the *Gibbs entropy* in the context of statistical mechanics.

Finally, for a Bayesian, this principle of indifference offers a natural way to pick a set of prior beliefs, and a prior distribution that respects this indifference principle is sometimes called a *noninformative prior*. While motivated in part by the Occam's razor heuristic, a Bayesian's *subjective* adoption of such a prior is very different from the automatic and objective embodiment of Occam's razor by the Bayes' factor, cf. our discussion of Bayesian model comparison in §6.2.2.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

Another important property of the entropy is its *additivity* for two random variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ that are described by a factorized joint distribution, $p(x, y) = p(x)\, p(y)$, and thus are statistically independent (1.79):

$$\begin{aligned}\mathcal{S}[p(x,y)] &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log p(x,y) \\ &= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x)p(y)\left[\log p(x) + \log p(y)\right] \\ &= -\sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{y\in\mathcal{Y}} p(y)\log p(y) \\ &= \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\,. \end{aligned} \tag{A.7}$$

Intuitively, if two observables are independent, then the expected amount of information learned from making an observation of each is just the sum of the information learned from making each individual observation. For macroscopic systems with small probabilities for individual outcomes, this makes the entropy a practically useful quantity to work with: while the probability of independent outcomes multiply and create smaller and smaller numbers, the surprisals, and thus the entropies, simply add.

The additivity property (A.7) has a very physical interpretation: the entropy is typically an **extensive** quantity, meaning that it scales with the number of **degrees of freedom** or microscopic size of the system. This should make sense given the counting-of-states interpretation of the entropy: if a variable x describes the potential contents of a 1 TB hard drive, and the variable y independently describes the potential contents of another 1 TB hard drive, then the total storage capacity of the hard drives together is additive and equal to 2 TB. As the number of states available to the combined "hard drive" system is the product of the states of the individual "hard drive" systems, their entropies are additive.

However, if the hard drives are not independent and are *constrained* so that one hard drive mirrors the other, then their total storage capacity is *subadditive* and instead would be equal to 1 TB.[3] In this case, the system had only half as many degrees of freedom as we naively thought it had due to the strict constraint creating strong correlations between the contents of the hard drives.

Thus, more generally for two statistically *dependent* observables constrained by a nonzero interaction between them, the entropy obeys a property called *subadditivity*:

$$\mathcal{S}[p(x,y)] < \mathcal{S}[p(x)] + \mathcal{S}[p(y)]\,. \tag{A.8}$$

In words, this means that the entropy of a joint distribution $p(x, y)$ will always be less than or equal to the sum of the entropies of the marginal distributions $p(x)$ and $p(y)$. This is also physically intuitive: if two random variables are statistically dependent,

[3] This configuration has a practical realization called *RAID 1*, where "RAID" stands for *Redundant Array of Independent Disks* and allows for fault tolerance by creating data redundancy between the two hard drives.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

then an observation of x conveys information about the likely outcome of making an observation of y, and so an observation of y doesn't convey as much information as it would have if we didn't already know x.[4]

[4]For the mathematically curious, we can turn this intuition into a quick proof. The usual route is to first prove the *Jensen inequality* and then apply it to an auxiliary object called the **Kullback-Leibler (KL) divergence**.

First let's state and then prove the Jensen inequality. Consider a discrete probability distribution over N elements $a_{\mu=1,\ldots,N}$ with corresponding probabilities p_μ such that $\sum_{\mu=1}^{N} p_\mu = 1$. Next, consider a convex function $f(a)$, i.e., a function that satisfies

$$f\big(\lambda a_1 + (1-\lambda)a_2\big) \geq \lambda f(a_1) + (1-\lambda) f(a_2), \tag{A.9}$$

for any $\lambda \in [0,1]$ and for any numbers a_1 and a_2 in the domain of the function. The Jensen inequality states that the expectation of such a function is greater than or equal to the function applied to the mean:

$$\mathbb{E}\left[f(a)\right] \geq f(\mathbb{E}\left[a\right]). \tag{A.10}$$

This can be proved by induction on N as follows. First, note that (A.10) holds trivially when $N = 1$. Then, see that

$$\begin{aligned}
\mathbb{E}\left[f(a)\right] \equiv \sum_{\mu=1}^{N+1} p_\mu f(a_\mu) &= p_{N+1} f(a_{N+1}) + (1-p_{N+1}) \sum_{\mu=1}^{N} \frac{p_\mu}{1-p_{N+1}} f(a_\mu) \\
&\geq p_{N+1} f(a_{N+1}) + (1-p_{N+1})\, f\left(\sum_{\mu=1}^{N} \frac{p_\mu}{1-p_{N+1}} a_\mu\right) \\
&\geq f\left(p_{N+1} a_{N+1} + (1-p_{N+1}) \sum_{\mu=1}^{N} \frac{p_\mu}{1-p_{N+1}} a_\mu\right) = f(\mathbb{E}\left[a\right]),
\end{aligned} \tag{A.11}$$

where in going from the first line to the second line we used the Jensen inequality (A.10) for N elements, and in going from the second line to the third line we used the convexity of the function (A.9).

As the next step, let us introduce the KL divergence (sometimes known as the *relative entropy*) between two *different* probability distributions $p(x)$ and $q(x)$ for a discrete variable, $x \in \mathcal{X}$,

$$KL\left[p(x)\,||\,q(x)\right] \equiv \sum_{x\in\mathcal{X}} p(x) \log\left[\frac{p(x)}{q(x)}\right] = -\mathcal{S}[p(x)] + \mathcal{S}[p(x), q(x)], \tag{A.12}$$

which is a multi-function functional of both distributions, $p(x)$ and $q(x)$, and importantly is *not* symmetric in its function arguments. Here $\mathcal{S}[p(x), q(x)] \equiv -\sum_{x\in\mathcal{X}} p(x) \log q(x)$ is an asymmetric quantity known as the *cross entropy*. As first mentioned in footnote 16 of §10.2.3, the KL divergence is an asymmetric measure of the closeness of two different distributions and is closely related to the *cross-entropy loss* (10.36).

Finally, let's show that the KL divergence is nonnegative by applying the Jensen inequality to the convex function $f(a) = -\log(a)$ with discrete elements $a_\mu = q(x)/p(x)$:

$$KL\left[p(x)\,||\,q(x)\right] \equiv \sum_{x\in\mathcal{X}} p(x) \log\left[\frac{p(x)}{q(x)}\right] \geq -\log\left[\sum_{x\in\mathcal{X}} p(x) \frac{q(x)}{p(x)}\right] = -\log(1) = 0. \tag{A.13}$$

To complete our proof, note that the positivity of the KL divergence (A.13) implies the subadditivty of the entropy (A.8) for a choice of distributions as $KL\left[p(x,y)\,||\,p(x)\,p(y)\right]$.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

Turning this argument upside down and shuffling (A.8) leftside right, let us define the **mutual information** between two random variables as:

$$\begin{aligned}\mathcal{I}\left[p(x,y)\right] &\equiv \mathcal{S}[p(x)] + \mathcal{S}[p(y)] - \mathcal{S}[p(x,y)] \\ &= \sum_{x\in\mathcal{X},y\in\mathcal{Y}} p(x,y)\log\left[\frac{p(x,y)}{p(x)\,p(y)}\right].\end{aligned} \tag{A.14}$$

This is a functional of a joint probability distribution and gives an average measure of how much information an observation of $x \in \mathcal{X}$ conveys about an observation of $y \in \mathcal{Y}$, and vice versa. Rearranged in this way, we see that the subadditivity of the entropy (A.8) implies the nonnegativity of the mutual information,

$$\mathcal{I}\left[p(x,y)\right] \geq 0, \tag{A.15}$$

with equality holding when and only when the sets of observable outcomes, $\mathcal{X}$ and $\mathcal{Y}$, are statistically independent.[5] Thus, the mutual information of a joint distribution is telling us something about the *interactions* that create the nontrivial correlations that are a signature of statistical dependence. We'll compute explicitly the mutual information of preactivations for infinite-width neural networks in §A.2 and for finite-width neural networks in §A.3: as you might imagine, the interaction of neurons at finite width will lead to nonzero mutual information.

As the last preparation before such computations, we will need to extend the definition of the entropy (A.1) from discrete outcomes to continuous random variables as

$$\mathcal{S}[p(x)] \equiv -\int dx\; p(x)\;\log p(x) = \mathbb{E}\left[s(x)\right]. \tag{A.16}$$

While the entropy is still the expectation of the surprisal (A.4), to take that expectation we now have to evaluate an integral rather than compute a sum. As passing to the continuum actually involves a physically interesting subtlety, let's take a few paragraphs to unravel this definition (A.16).

First, let's understand this subtlety mathematically. Since our random variable is continuous, we can make a smooth and monotonic change of coordinates in our continuous *outcome space* $\mathcal{X}$ from x to $\widetilde{x} \equiv \widetilde{x}(x)$. Since such coordinates are arbitrary, consistency requires that the probability of observing x in the interval between $[x_1, x_2]$ be the same as the probability of observing $\widetilde{x}$ in $[\widetilde{x}(x_1), \widetilde{x}(x_2)]$. In equations, this means that

$$p(x_1 < x < x_2) \equiv \int_{x_1}^{x_2} dx\; p(x) \tag{A.17}$$

[5] Note that the mutual information (A.14) can also be thought of as the KL divergence (A.12) between the joint distribution $p(x,y)$ and the product of the marginal distributions $p(x)\,p(y)$. That is, the mutual information $\mathcal{I}\left[p(x,y)\right]$ tells us how close a joint distribution is to being a product of independent distributions, with nonzero mutual information signaling statistical dependence.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

must equal

$$p(\widetilde{x}_1 < \widetilde{x} < \widetilde{x}_2) \equiv \int_{\widetilde{x}(x_1)}^{\widetilde{x}(x_2)} d\widetilde{x}\; p(\widetilde{x}) = \int_{x_1}^{x_2} dx\; \frac{d\widetilde{x}}{dx}\; p\Big(\widetilde{x}(x)\Big), \tag{A.18}$$

where in the last step we used the standard transformation property of the integral measure under a change of coordinates. Thus, the two distributions $p(\widetilde{x})$ and $p(x)$ must be related as

$$p\Big(\widetilde{x}(x)\Big) \equiv \frac{dx}{d\widetilde{x}}\; p(x), \tag{A.19}$$

which is the well-known coordinate transformation formula for a probability density of a single variable. What about the expected surprisal? Under this same coordinate transformation, the entropy (A.16) is given by

$$\begin{aligned} \mathcal{S}[p(\widetilde{x})] &= -\int d\widetilde{x}\; p(\widetilde{x})\; \log p(\widetilde{x}) \\ &= -\int dx\; \frac{d\widetilde{x}}{dx}\; \frac{dx}{d\widetilde{x}}\; p(x)\; \left[\log p(x) + \log\left(\frac{dx}{d\widetilde{x}}\right)\right] \\ &= \mathcal{S}[p(x)] + \int dx\; p(x)\; \log\left(\frac{d\widetilde{x}}{dx}\right)\,, \end{aligned} \tag{A.20}$$

where in the second line we again used the standard transformation property of the integral measure under a change of coordinates, and we also used (A.19) to express the probability distribution in the old coordinates. So with a general (monotonic) change of coordinates, we can make the entropy take any value we'd like by a judicious choice of the Jacobian of the transformation, $d\widetilde{x}/dx$ – even a negative one!

Now, let's understand the physical meaning of this subtlety. Let's consider a coordinate change that's just a multiplicative factor, $\widetilde{x} = cx$, which is like changing the base unit that we use to measure the observable x from `inches` to `meters`. In this case, the surprisal of each particular outcome shifts by the same constant, $\Delta s(x) = \log c$, and so does the entropy. Thus, for continuous quantities the entropy is additively sensitive to the choice of units with which we measure our observable quantities, and we really should specify these units along with the definition of the entropy (A.16).[6]

[6] For any *dimensionful* observable, the probability *density* $p(x)$ must also be dimensionful so that the probability of observing x in the interval $[x_1, x_2]$, $p(x_1 < x < x_2) \equiv \int_{x_1}^{x_2} dx\; p(x)$, is properly dimensionless. Yet, putting such a dimensionful object $p(x)$ into an argument of the logarithm as $\log p(x)$ is illegitimate, as we discussed in footnote 15 of §1. This is another way of seeing that for continuous probability distributions, i.e., probability densities, we need to specify the measuring units to properly define the entropy.

For the curious and potentially worried reader, it should be noted here that the *Bayes' factor* (6.30) that contained the observation dependence of our Bayesian model comparison is invariant under a coordinate transformation of our observations $y_{\mathcal{A}}$. What ultimately mattered there was the relative magnitude of the *evidence* of different hypotheses and not the absolute value of the individual evidences.

Also, please do not confuse this issue of the units for a probability density, which change the entropy by an additive constant, with the units of entropy that we discussed in footnote 1, which change the entropy by a multiplicative constant. Note that for discrete probability distributions, the probability distribution gives a simple probability and is already dimensionless, cf. our brief discussion in footnote 1.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

Perhaps the most sensible choice is to set the measuring units according to the smallest possible measurements of x that we can physically distinguish – in other words, according to the precision limit set by the constraints of the physical world. This precision limit ϵ is sometimes called a *cutoff* and in practice means that we only care about the discrete probability of finding x between $[x_0, x_0+\epsilon]$.[7] Such a discretization of the outcome space will then ensure that the entropy is always positive, (A.6), since we're now dealing with discrete probabilities again.[8]

If this physical sensitivity of the entropy to the somewhat arbitrary cutoff bothers you, then perhaps you should consider the continuous analog of mutual information,

$$\begin{aligned} \mathcal{I}\left[p(x,y)\right] &\equiv \mathcal{S}[p(x)] + \mathcal{S}[p(y)] - \mathcal{S}[p(x,y)] \\ &= \int dxdy\; p(x,y)\, \log\left[\frac{p(x,y)}{p(x)\,p(y)}\right], \end{aligned} \tag{A.21}$$

where the definition in terms of the entropy is the same as in the discrete case, (A.14). In particular, mutual information is insensitive to the choice of the measuring coordinates. To see why, let's consider two continuous random variables x and y and independent monotonic coordinate transformations,

$$p\Big(\widetilde{x}(x)\Big) = \frac{dx}{d\widetilde{x}}\; p(x)\,, \qquad p\Big(\widetilde{y}(y)\Big) = \frac{dy}{d\widetilde{y}}\; p(y)\,, \qquad p\Big(\widetilde{x}(x),\widetilde{y}(y)\Big) = \frac{dx}{d\widetilde{x}}\frac{dy}{d\widetilde{y}}\; p(x,y), \tag{A.22}$$

where to get this we applied a similar consistency-of-probabilities argument to the one that we gave above. With this in mind, we can now show that the mutual information (A.21) stays invariant under these coordinate transformations:

$$\begin{aligned} \mathcal{I}\left[p(\widetilde{x},\widetilde{y})\right] &\equiv \mathcal{S}[p(\widetilde{x})] + \mathcal{S}[p(\widetilde{y})] - \mathcal{S}[p(\widetilde{x},\widetilde{y})] \\ &= \int d\widetilde{x}d\widetilde{y}\; p(\widetilde{x},\widetilde{y})\, \log\left[\frac{p(\widetilde{x},\widetilde{y})}{p(\widetilde{x})\,p(\widetilde{y})}\right] \\ &= \int dxdy\; p(x,y)\, \log\left[\frac{p(x,y)}{p(x)\,p(y)}\right] \equiv \mathcal{I}\left[p(x,y)\right]. \end{aligned} \tag{A.23}$$

In going from the second line to the third line, the coordinate transformation factors $dx/d\widetilde{x}$ and $dy/d\widetilde{y}$ completely cancelled inside the logarithm, and the transformation

[7] Please don't confuse our *measurement precision cutoff* here with the *perturbative cutoff* of the effective theory, the depth-to-width ratio $r \equiv L/n$. While in both cases we can think of them as important scales, they have very different physical meanings: in the former case, the precision cutoff gives the minimum distinguishable difference between measurements of two observables, $\epsilon \equiv \min(|z_2 - z_1|)$; in the latter case, the cutoff of the effective theory, L/n, sets the scale at which finite-width corrections need to be taken into account in the preactivation distribution $p(z)$.

[8] In the context of deep learning, a natural choice is the precision limit of the floating-point representation of the network's variables. Since type `float` eponymously has a precision that's relative to the value being stored, one could perhaps choose the minimum precision in the relevant range.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

of the measure cancelled the coordinate transformation factors outside the logarithm. Thus, the mutual information is completely well defined for continuous random variables, independent of the cutoff ϵ. For this reason, with a consistent and fixed choice of units, it's completely valid to compute the entropy as an intermediate step in the computation of the mutual information, and we will make use of this fact in the following sections.

Finally, note that the notion of the mutual information can be extended to more than two random variables. For instance, in §A.3 we will consider the **tripartite information** among three random variables $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and $z \in \mathcal{Z}$:

$$\begin{aligned}
&\mathcal{I}_3\left[p(x,y,z)\right] \qquad\qquad (A.24)\\
&\quad \equiv \mathcal{I}\left[p(x,y)\right] - \mathcal{I}\left[p(x,y|z)\right]\\
&\quad = \mathcal{S}[p(x)] + \mathcal{S}[p(y)] + \mathcal{S}[p(z)] - \mathcal{S}[p(x,y)] - \mathcal{S}[p(y,z)] - \mathcal{S}[p(z,x)] + \mathcal{S}[p(x,y,z)]\\
&\quad = \sum_{x\in\mathcal{X},y\in\mathcal{Y},z\in\mathcal{Z}} p(x,y,z)\,\log\left[\frac{p(x,y)\,p(y,z)\,p(z,x)}{p(x)\,p(y)\,p(z)\,p(x,y,z)}\right].
\end{aligned}$$

Here, $\mathcal{I}\left[p(x,y|z)\right]$ is the mutual information of the joint distribution between x and y conditioned on z, $p(x,y|z)$, and the final expression for $\mathcal{I}_3\left[p(x,y,z)\right]$ makes it clear that *(a)* it is fully symmetric in its three arguments and that *(b)* its continuous analog that's in your imagination is cutoff-independent and invariant under similar coordinate transformations as those in (A.22).

What is not immediately obvious is that tripartite information can be either positive or negative. From the first expression in (A.24), we can gain some intuition for the meaning of the tripartite information: it is a measure of whether knowledge of a third random variable, z in this way of writing the expression, increases or decreases the mutual information between the other two variables. When positive, it indicates that z contains information about x and y, and so knowing z decreases the amount of information you'd learn about x by measuring y; the information is stored *redundantly* between these three variables. When negative, it indicates that the information is distributed across x, y, and z in such a way that you'd learn less about x by measuring y than you would with first knowing z; in this case, the information is stored *synergistically* between these three variables.[9]

[9] An extreme example of positive tripartite information occurs when x, y, and z are completely dependent and exact copies of each other. In this redundant case, $\mathcal{I}\left[p(x,y)\right] > 0$, since knowledge of y tells you everything about x, but $\mathcal{I}\left[p(x,y|z)\right] = 0$, since conditioning on z means that there's nothing left for you to learn about x by also observing y. An example of such a situation would be three copies of the same book.

An extreme example of negative tripartite information occurs when the joint distribution between x and y factorizes, $p(x,y) = p(x)\,p(y)$, but the joint distribution conditioned on z does not, $p(x,y|z) \neq p(x|z)\,p(y|z)$. In this synergistic case, $\mathcal{I}\left[p(x,y)\right] = 0$, since without z the variables are statistically independent, but $\mathcal{I}\left[p(x,y|z)\right] > 0$, since there are correlations that are mediated by z. An example of such a situation could be a code that distributes a key among three different parties: knowledge of any two parts would give absolutely no information about the key, but with all three parts together the key can be reconstructed.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

A.2 Information at Infinite Width: Criticality

With that informative introduction out of the way, let's now make these abstract definitions concrete by using them to better understand the neural-network prior distribution.

To begin, let's focus on m preactivations $z_{i;\alpha}^{(\ell)}$ from the ℓ-th layer of an infinite-width neural network. Depending on when you're coming to this appendix from the main text, it's probably ingrained in your mind already that such preactivations are distributed according to a zero-mean Gaussian distribution

$$p\Big(z_1,\ldots,z_m\Big|\mathcal{D}\Big) = \frac{1}{\sqrt{|2\pi K|^m}} \exp\left(-\frac{1}{2}\sum_{i=1}^{m}\sum_{\alpha_1,\alpha_2\in\mathcal{D}} K^{\alpha_1\alpha_2} z_{i;\alpha_1} z_{i;\alpha_2}\right), \tag{A.25}$$

where in this expression, and in this section, we will drop layer indices everywhere to declutter expressions. The entropy (A.16) of this distribution is then given by

$$\begin{aligned}\mathcal{S}\Big[p\Big(z_1,\ldots,z_m\Big|\mathcal{D}\Big)\Big] &= \left\langle\!\!\left\langle \frac{1}{2}\sum_{i=1}^{m}\sum_{\alpha_1,\alpha_2\in\mathcal{D}} K^{\alpha_1\alpha_2} z_{i;\alpha_1} z_{i;\alpha_2} \right\rangle\!\!\right\rangle_K + \log\left(\sqrt{|2\pi K|^m}\right) \\ &= \frac{m}{2}\left(N_{\mathcal{D}} + \log|2\pi K|\right) = \frac{m}{2}\log|2e\pi K| \ ,\end{aligned} \tag{A.26}$$

where as a reminder $|2\pi eK|$ is the determinant of the $N_{\mathcal{D}}$-by-$N_{\mathcal{D}}$ matrix $2\pi e K_{\alpha_1\alpha_2}$.[10] From this we conclude that entropy is *extensive*, proportional to the number of neurons m in the joint distribution (A.25). This shows how the entropy can count the *degrees of freedom* in a deep learning context – in this case, by counting the neurons – and the exact additivity in the number of neurons signals to us that the individual neurons in an infinite-width layer are noninteracting and statistically independent.

To confirm this directly, let's work out the mutual information between two sets of neurons, $\mathcal{M}_1 = \{1,\ldots,m_1\}$ and $\mathcal{M}_2 = \{m_1+1,\ldots,m_1+m_2\}$, in the same layer ℓ. In excruciating detail for its triviality, we have

$$\begin{aligned}&\mathcal{I}\left[p\Big(\mathcal{M}_1,\mathcal{M}_2\Big|\mathcal{D}\Big)\right] \\ &\quad= \mathcal{S}\Big[p\Big(z_1,\ldots,z_{m_1}\Big|\mathcal{D}\Big)\Big] + \mathcal{S}\Big[p\Big(z_{m_1+1},\ldots,z_{m_1+m_2}\Big|\mathcal{D}\Big)\Big] - \mathcal{S}\Big[p\Big(z_1,\ldots,z_{m_1+m_2}\Big|\mathcal{D}\Big)\Big] \\ &\quad= \left[m_1+m_2-(m_1+m_2)\right]\frac{1}{2}\log|2e\pi K| = 0,\end{aligned} \tag{A.27}$$

where to go to the last line we used (A.26) three different ways. This zero mutual information confirms that at infinite width, learning the activities of some neurons in a

[10] In this and the next section, for convenience we will ignore the ambiguity in the definition of the continuous entropy and use the formula (A.16) naively. This is permissible since we ultimately are interested in cutoff-independent quantities, such as the mutual information, and when interpreting an entropy such as (A.26), we will never care about its absolute value.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

layer conveys no information about the activities of any of the other neurons. To find a finite result, we'll have to back off the infinite-width limit (§A.3).

That said, for a fixed neuron we do expect nontrivial correlations between different inputs and therefore also finite mutual information. To investigate this, let's take two inputs $x_{i;+}$ and $x_{i;-}$ and compute the mutual information between the preactivations for a particular neuron, $z_{1;+}$ and $z_{1;-}$. Plugging the entropy (A.26) into the definition of the mutual information (A.14), we find

$$\begin{aligned} \mathcal{I}\left[p(z_{1;+}, z_{1;-}|x_\pm)\right] &= \frac{1}{2}\left[\log(K_{++}) + \log(K_{--}) - \log\Big(K_{++}K_{--} - K_{+-}^2\Big)\right] \\ &= \frac{1}{2}\log\left(\frac{K_{++}K_{--}}{K_{++}K_{--} - K_{+-}^2}\right). \end{aligned} \tag{A.28}$$

Focusing on inputs $x_{i;\pm}$ with equal norms such that $K_{++} = K_{--} = K_{[0]} + K_{[2]}$ and $K_{+-} = K_{[0]} - K_{[2]}$, cf. (5.17) and (5.19), we can rewrite this mutual information as

$$\mathcal{I}\left[p(z_{1;+}, z_{1;-})\right] = \frac{1}{2}\log\left[\frac{\left(1 + \frac{K_{[2]}}{K_{[0]}}\right)^2}{4\frac{K_{[2]}}{K_{[0]}}}\right], \tag{A.29}$$

which is parameterized entirely by the dimensionless ratio $K_{[2]}/K_{[0]}$ that for nearby inputs $x_{i;\pm} = x_{i;0} \pm \delta x_i/2$ characterizes their relative angle after passing through the network to the ℓ-th layer.

There are two interesting limits here. First, when $K_{[2]}/K_{[0]} \to 0$, the mutual information becomes large. This follows because as the relative angle between the preactivations vanishes, they become close to each other: knowing the preactivation for one input tells us about the value of the preactivation for the other input. In a classification problem, this is a great prior if the inputs are from the same class but would make learning really difficult if they're from different classes. Second, when $K_{[0]} = K_{[2]}$, the mutual information vanishes. This follows because in this limit the off-diagonal components of the kernel, $K_{+-} = K_{[0]} - K_{[2]}$, vanish: the preactivations become statistically independent. In a classification problem, this may be a good prior if the inputs are from different classes – so long as the details of the classes don't correlate in some way – but would make learning really difficult if the inputs are from the same class. Altogether, this suggests that as a prior we don't want $K_{[2]}/K_{[0]}$ to be too big or too small, which can be best ensured by setting both $\chi_\parallel = 1$ and $\chi_\perp = 1$, cf. (5.55). This gives an information-theoretic perspective on *criticality*.

Finally, while we have focused here on the prior distribution for networks at initialization, we could also study these same quantities after learning using the Bayesian posterior (6.66) or the fully-trained distribution of gradient-based learning (10.39). Since the entropy of a Gaussian distribution is independent of its mean – it can be eliminated by a change of dummy integration variables – the mutual information for either trained infinite-width network is given by this same expression, (A.28), but with the kernel replaced by the covariance of the Bayesian posterior distribution (6.57) or the covariance

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

of the generalized posterior distribution (10.41). In the latter case, the mutual information will involve both the kernel and the frozen NTK, and its analysis would yield similar results to those that we found in §10.3.1 when we investigated the bias–variance tradeoff. This would give an information-theoretic perspective on generalization.[11]

A.3 Information at Finite Width: Optimal Aspect Ratio

In this section, we'll see how finite-width networks have a prior for nonzero mutual information between different neurons. Assuming that nonzero mutual information is desirable by intuition – and by analogy to an *unsupervised* learning objective – we can use this computation to optimize the depth-to-width ratio for finite-width MLPs at criticality. This *optimal aspect ratio*, $r^\star$, defines the scale that separates effectively-deep networks – describable by our effective theory approach – from overly-deep networks – not describable due to strong interactions and not trainable due to large fluctuations.[12]

In general, the entropy and mutual information of any interacting theory are really difficult to compute. However, when the interactions are *nearly-Gaussian*, we can use perturbation theory. To do so, there is actually a neat *variational principle* that we can use to organize our calculations, so let us explain that first.

As in the last section, we'll focus on the distribution of m preactivations in layer ℓ, drop (almost) all the layer indices, and focus exclusively on a single input x. As we have been doing since the beginning of time (§1.3), let's express the probability distribution in terms of an action as

$$p(z_1, \ldots, z_m|x) = \frac{e^{-S(z_1,\ldots,z_m)}}{Z}, \tag{A.30}$$

which must be normalized by the *partition function*

$$Z = \int \left[\prod_{i=1}^{m} dz_i\right] e^{-S(z_1,\ldots,z_m)}. \tag{A.31}$$

In terms of these quantities, the entropy (A.16) can be expressed as

$$\mathcal{S}[p(z_1, \ldots, z_m|x)] = \log(Z) + \frac{1}{Z}\int \left[\prod_{i=1}^{m} dz_i\right] e^{-S(z_1,\ldots,z_m)} S(z_1, \ldots, z_m). \tag{A.32}$$

Just to make sure, please don't get confused between the action $S(z)$, which is a function of the preactivations, and the entropy $\mathcal{S}[p(z)]$, which is a functional of the probability distribution.[13]

[11] Analogously, studying the tripartite information generalization of (A.28) for either type of posterior distribution would give an alternative perspective on the $*$-polation results of §10.3.2.

[12] Note that in this section, we'll need the expressions for the running couplings that we derived in §4.4 when finding the effective distribution of m neurons in a wide-but-finite layer. As such, it may be helpful to reread that section before proceeding.

[13] For choices of units of the preactivations z for which the partition function is unity, $Z = 1$, the action is simply the surprisal (A.4), and thus the entropy is the expectation of the action: $\mathcal{S}[p(z)] = \mathbb{E}\left[S(z)\right]$.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

To proceed, let's adopt a **variational ansatz**: we'll divide the action into two parts as

$$S(z_1, \ldots, z_m) = S_{\mathrm{F}}(z_1, \ldots, z_m) + S_{\mathrm{I}}(z_1, \ldots, z_m), \tag{A.33}$$

with the idea being that the second term, S_{I}, encodes the part of the distribution that is perturbatively small. Specifically, the **variational principle** instructs us to choose the first term in (A.33), $S_{\mathrm{F}}(z)$, that gives no variations of the entropy with respect to $S_{\mathrm{I}}(z)$:

$$0 = \left.\frac{\delta \mathcal{S}[p(z)]}{\delta S_{\mathrm{I}}(z)}\right|_{S_{\mathrm{I}}(z)=0}. \tag{A.34}$$

We'll satisfy this shortly. Additionally, $S_{\mathrm{I}}(z)$ should not be completely arbitrary, but should instead be constructed to properly reflect the statistics of the preactivation distribution $p(z_1, \ldots, z_m|x)$. The first such constraint comes from demanding that the two-point correlator, when computed with the variational action, is determined by the *exact* single-input metric G:

$$\mathbb{E}\left[z_{i_1} z_{i_2}\right] \equiv \delta_{i_1 i_2} G. \tag{A.35}$$

The second constraint comes from demanding that the connected four-point correlator, when computed with the variational action, is determined by the *exact* single-input four-point vertex:

$$\mathbb{E}\left[z_{i_1} z_{i_2} z_{i_3} z_{i_4}\right]|_{\text{connected}} = \frac{1}{n}\left(\delta_{i_1 i_2}\delta_{i_3 i_4} + \delta_{i_1 i_3}\delta_{i_2 i_4} + \delta_{i_1 i_4}\delta_{i_2 i_3}\right) V. \tag{A.36}$$

Together, the constraints (A.35) and (A.36) will fix the couplings of the variation action S_{I} so that the full action (A.33) correctly specifies the m-neuron preactivation distribution (A.30).[14]

To understand why we're doing this, note that our variational principle is ultimately just a realization of the **principle of maximum entropy** for nearly-Gaussian distributions.[15] In particular, first note that the Gaussian distribution itself can be derived

Note that in thermodynamics, the constant $-\log Z$ is sometimes called the *free energy*; this discussion should make clear that only its relative value for two different distributions is physical.

[14] In principle there are additional constraints coming from the statistics of higher-point correlators, but their contribution is subleading to both the leading and next-to-leading orders that we will work with.

[15] For those readers that enjoy our historical asides, the maximum entropy principle was discovered by Jaynes and provides a link between *information theory* on the one hand and *statistical mechanics* on the other hand [80, 81]. As an example of this, consider a central problem in statistical mechanics: find the probability distribution p_i for the fundamental *microstates* i of a system that has a macroscopic average energy $\overline{E} \equiv \mathbb{E}\left[E_i\right] = \sum_i p_i E_i$. An application of the principle of maximum entropy then correctly picks out the *Boltzmann distribution* (or sometimes, the *Gibbs distribution*) of statistical mechanics,

$$p_i \propto e^{-\beta E_i}, \tag{A.37}$$

if we optimize the entropy (A.1) subject to the observable constraint for the energy, $\sum_i p_i E_i = \overline{E}$, and the normalization condition for the distribution, $\sum_i p_i = 1$. Here, β is a Lagrange multiplier that depends

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

as a distribution that maximizes the entropy (A.16), subject to the constraints of fixed first and second moments.[16] As we will see in a moment, in (A.34) we are maximizing the entropy of the distribution with respect to the deformation of the action away from Gaussianity, $S_{\rm I}$, subject to the constraint of fixing the higher-order cumulants order-by-order in perturbation theory. In general, the maximum entropy principle is an appropriate procedure when we have fixed observable information for which we want to find an underlying distribution. Here, we actually know the distribution that produces G and V, (A.30), but we can still use this principle as convenient tool for computing the entropy.[17]

Now, to satisfy the variational principle (A.34), let's choose

$$S_{\rm F}(z_1, \ldots, z_m) = \frac{1}{2G} \sum_{i=1}^{m} z_i^2 . \tag{A.38}$$

Importantly, G is not just the inverse of the quadratic coefficient in the action $S(z)$ but instead is the *exact* two-point correlator that we would actually measure, (A.35), incorporating the full series of corrections due to the interactions, cf. (4.104).[18] To see why such a choice satisfies the variational principle, let us start by rewriting expectations with respect to the full distribution (A.30) in terms of simpler *Gaussian* expectations taken with respect to a zero-mean Gaussian distribution with the same variance (A.35):

$$\langle\!\langle z_{i_1} z_{i_2} \rangle\!\rangle_G = \delta_{i_1 i_2} G . \tag{A.39}$$

Here, please recall our notation $\langle\!\langle \cdot \rangle\!\rangle_G$ for a Gaussian expectation of a multi-neuron function with variance $\delta_{i_1 i_2} G$,

$$\langle\!\langle f(z_1, \ldots, z_m) \rangle\!\rangle_G \equiv \frac{1}{Z_G} \int \left[\prod_{i=1}^{m} dz_i \right] e^{-\frac{1}{2G} \sum_{i=1}^{m} z_i^2} f(z_1, \ldots, z_m), \tag{A.40}$$

and note also that such a Gaussian distribution will require a different partition function

$$Z_G \equiv \int \left[\prod_{i=1}^{m} dz_i \right] e^{-\frac{1}{2G} \sum_{i=1}^{m} z_i^2} = (2\pi G)^{\frac{m}{2}} \; ; \tag{A.41}$$

on the energy $\overline{E}$ and has a physical interpretation as the inverse temperature, $\beta = 1/(k_{\rm B} T)$, with the aforementioned Boltzmann constant $k_{\rm B}$ and the familiar-to-everyone temperature T. This example also shows how statistical mechanics links the details of the fundamental microstates i to the macroscopic thermodynamic variables such as $\overline{E}$ and T.

[16] For those of you keeping track: the maximum entropy distribution with no information fixed is the uniform distribution, cf. (A.3); the maximum entropy distribution with a fixed first moment is the Boltzmann distribution, cf. (A.37); and the maximum entropy distribution with fixed first and second moments is the Gaussian distribution, cf. (nearly-)everywhere.

[17] In particular, this procedure will organize our perturbative computation of the entropy and ensure that we only need to compute Gaussian expectations of the form (A.40).

[18] This will let us express the entropy in terms of these measurable quantities, and in no way will we need to actually compute any of the corrections in this series.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

importantly, $Z \neq Z_G$. Next, let us rewrite the entropy (A.32) in terms of these simpler expectations as

$$\begin{aligned}\mathcal{S}[p(z_1,\ldots,z_m|x)] &= \log Z + \mathbb{E}\left[\frac{1}{2G}\sum_{i=1}^m z_i^2\right] + \frac{1}{Z}\int\left[\prod_{i=1}^m dz_i\right] e^{-\frac{1}{2G}\sum_{i=1}^m z_i^2} e^{-S_{\mathrm{I}}} S_{\mathrm{I}} \\ &= \frac{m}{2}\log(2\pi G) + \log\left(\frac{Z}{Z_G}\right) + \frac{m}{2} + \left(\frac{Z}{Z_G}\right)^{-1} \left\langle\!\!\left\langle e^{-S_{\mathrm{I}}} S_{\mathrm{I}} \right\rangle\!\!\right\rangle_G , \end{aligned} \tag{A.42}$$

where in the first equality we just plugged in (A.33), and in the second equality we rewrote all the full expectations in terms of the simpler Gaussian expectations using (A.40). Then, by Taylor-expanding in S_{I}, we can evaluate the ratio of partition functions as

$$\begin{aligned}\frac{Z}{Z_G} &= \frac{1}{Z_G}\int\left[\prod_{i=1}^m dz_i\right] e^{-\frac{1}{2G}\sum_{i=1}^m z_i^2}\left(e^{-S_{\mathrm{I}}}\right) = \left\langle\!\!\left\langle e^{-S_{\mathrm{I}}}\right\rangle\!\!\right\rangle_G \\ &= 1 - \langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G + \frac{1}{2}\left\langle\!\!\left\langle S_{\mathrm{I}}^2\right\rangle\!\!\right\rangle_G - \frac{1}{6}\left\langle\!\!\left\langle S_{\mathrm{I}}^3\right\rangle\!\!\right\rangle_G + O\left(S_{\mathrm{I}}^4\right) , \end{aligned} \tag{A.43}$$

and similarly we can evaluate the needed Gaussian expectation as

$$\left\langle\!\!\left\langle e^{-S_{\mathrm{I}}} S_{\mathrm{I}}\right\rangle\!\!\right\rangle_G = \langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G - \left\langle\!\!\left\langle S_{\mathrm{I}}^2\right\rangle\!\!\right\rangle_G + \frac{1}{2}\left\langle\!\!\left\langle S_{\mathrm{I}}^3\right\rangle\!\!\right\rangle_G + O\left(S_{\mathrm{I}}^4\right). \tag{A.44}$$

Plugging these back into the variational expression for the entropy (A.42) and organizing terms, for which you might find $\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} + \cdots$ and $1/(1+x) = 1 - x + x^2 - x^3 + \cdots$ helpful, we get

$$\begin{aligned}\mathcal{S}[p(z_1,\ldots,z_m|x)] &= \frac{m}{2}\log(2\pi e G) - \frac{1}{2}\left[\left\langle\!\!\left\langle S_{\mathrm{I}}^2\right\rangle\!\!\right\rangle_G - \langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G^2\right] \\ &\quad + \frac{1}{3}\left[\left\langle\!\!\left\langle S_{\mathrm{I}}^3\right\rangle\!\!\right\rangle_G - 3\left\langle\!\!\left\langle S_{\mathrm{I}}^2\right\rangle\!\!\right\rangle_G \langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G + 2\langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G^3\right] + O\left(S_{\mathrm{I}}^4\right). \end{aligned} \tag{A.45}$$

First, note that the first term is exactly the entropy for a multivariate Gaussian distribution with a covariance $\delta_{i_1 i_2} G$, cf. (A.26). Second, and most importantly, note that the would-be linear term proportional to $\langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G$ exactly cancelled out. In other words, our ansatz for the decomposition of the action (A.33) with the choice (A.38) automatically satisfies the variational principle (A.34).

Finally, let us note in passing that the leading correction coming from the quadratic term is definitively negative.[19] This negativity is actually necessary for mathematical consistency: as the Gaussian distribution maximizes the entropy of any set of random variables with known and fixed first and second moments, any deformation of a distribution away from Gaussianity, while also respecting such moment constraints, must

[19] To see this, notice that the expression in the square brackets of the quadratic term is the variance of the variational part of the action, and thus is positive: $\left\langle\!\!\left\langle S_{\mathrm{I}}^2\right\rangle\!\!\right\rangle_G - \langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G^2 = \langle\!\langle \left(S_{\mathrm{I}} - \langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G\right)^2 \rangle\!\rangle_G \geq 0$.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

necessarily have less entropy. In the current case, our variational ansatz (A.33) gives a nearly-Gaussian deformation of a zero-mean Gaussian distribution.[20]

Now that we're done passing notes, let's satisfy our constraints, (A.35) and (A.36), and then use our variational expression (A.45) to compute the entropy and mutual information of the preactivations in a finite-width network.

Leading-Order Correction: Nonzero Mutual Information

At leading order, we've already worked out how to relate the couplings of the variational action $S_{\rm I}$ to the single-input metric G and the single-input four-point correlator V. Recall from our discussion of the running couplings when integrating out neurons in §4.4 that the leading correction to the action was given by

$$S_{\rm I}(z_1,\ldots,z_m) = \frac{1}{2}\left(g_m - \frac{1}{G}\right)\sum_{i=1}^{m} z_i^2 - \frac{v_m}{8}\sum_{i,j=1}^{m} z_i^2 z_j^2 + O\left(\frac{1}{n^2}\right), \tag{A.46}$$

where the running quadratic coupling was given by (4.102),

$$g_m = \frac{1}{G} + \left(\frac{m+2}{2}\right)\frac{V}{nG^3} + O\left(\frac{1}{n^2}\right), \tag{A.47}$$

and the running quartic coupling was given by (4.103),

$$v_m = \frac{V}{nG^4} + O\left(\frac{1}{n^2}\right). \tag{A.48}$$

To get the expression (A.46), look at our expression for the distribution of m preactivations, (4.97), and then rearrange (A.33) with (A.38) to solve for $S_{\rm I}$. If you don't recall how to get (A.47) and (A.48), feel free to flip back and reread the last subsubsection of §4.4, or feel free to flip forward to the next subsection where we'll have to derive these expressions again to higher order in the $1/n$ expansion, cf. (A.59) and (A.63).

Given that this leading-order variational correction to the action $S_{\rm I}$ (A.46) now satisfies the constraints (A.35) and (A.36), we can now evaluate the leading correction to the entropy (A.45):

$$\begin{aligned}
&\left\langle\!\left\langle S_{\rm I}^2 \right\rangle\!\right\rangle_G - \left\langle\!\left\langle S_{\rm I} \right\rangle\!\right\rangle_G^2 \\
&\quad = \frac{1}{4}\left(g_m - \frac{1}{G}\right)^2 \sum_{i_1,i_2=1}^{m} \left[\left\langle\!\left\langle z_{i_1}^2 z_{i_2}^2 \right\rangle\!\right\rangle_G - \left\langle\!\left\langle z_{i_1}^2 \right\rangle\!\right\rangle_G \left\langle\!\left\langle z_{i_2}^2 \right\rangle\!\right\rangle_G\right] \\
&\qquad - \frac{1}{8}\left(g_m - \frac{1}{G}\right) v_m \sum_{i_1,i_2,i_3=1}^{m} \left[\left\langle\!\left\langle z_{i_1}^2 z_{i_2}^2 z_{i_3}^2 \right\rangle\!\right\rangle_G - \left\langle\!\left\langle z_{i_1}^2 z_{i_2}^2 \right\rangle\!\right\rangle_G \left\langle\!\left\langle z_{i_3}^2 \right\rangle\!\right\rangle_G\right]
\end{aligned} \tag{A.49}$$

[20] N.B. the terms in the second set of the square brackets in (A.45) are necessary for computing the next-to-leading-order correction.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

$$
\begin{aligned}
&+\frac{1}{64}v_m^2\sum_{i_1,i_2,i_3,i_4=1}^{m}\left[\left\langle\!\left\langle z_{i_1}^2 z_{i_2}^2 z_{i_3}^2 z_{i_4}^2\right\rangle\!\right\rangle_G-\left\langle\!\left\langle z_{i_1}^2 z_{i_2}^2\right\rangle\!\right\rangle_G\left\langle\!\left\langle z_{i_3}^2 z_{i_4}^2\right\rangle\!\right\rangle_G\right]+O\left(\frac{1}{n^3}\right)\\
&=\frac{m}{2}\left(g_m-\frac{1}{G}\right)^2G^2-\frac{m(m+2)}{2}\left(g_m-\frac{1}{G}\right)v_mG^3+\frac{m(m+2)(m+3)}{8}v_m^2G^4\\
&\quad+O\left(\frac{1}{n^3}\right).
\end{aligned}
$$

Here, in going from the second line to the last line, you may find this formula for these Gaussian expectations useful:

$$
\sum_{i_1,\ldots,i_k=1}^{m}\left\langle\!\left\langle z_{i_1}^2\cdots z_{i_k}^2\right\rangle\!\right\rangle_G=m(m+2)\cdots[m+2(k-1)]\,G^k, \tag{A.50}
$$

which is akin to (3.46) and will be used again in the next subsubsection repeatedly, up to $k=6$. To complete the computation, plug in the quadratic coupling, (A.47), and the quartic coupling, (A.48), into (A.49), giving

$$
\left\langle\!\left\langle S_{\mathrm{I}}^2\right\rangle\!\right\rangle_G-\langle\!\langle S_{\mathrm{I}}\rangle\!\rangle_G^2=\frac{m(m+2)}{8}\left(\frac{V}{nG^2}\right)^2+O\left(\frac{1}{n^3}\right) \tag{A.51}
$$

for the variance of the variational action. Therefore, the entropy (A.45) is given by

$$
\mathcal{S}[p(z_1,\ldots,z_m|x)]=\frac{m}{2}\log(2\pi eG)-\frac{(m^2+2m)}{16}\left(\frac{V}{nG^2}\right)^2+O\left(\frac{1}{n^3}\right), \tag{A.52}
$$

exhibiting a nontrivial correction at finite width.

Let us reflect on this formula by making some comments. First, note that the correction is definitely negative, as we pointed out before: recall that the Gaussian distribution maximizes the entropy of a set of random variables with a fixed covariance, and since our first variational constraint (A.35) fixes the two-point correlator of the preactivations, the entropy for the nearly-Gaussian distribution (A.30) must be less than the entropy of a Gaussian distribution with the same variance. Second, note that unlike all our previous results, the leading correction here is *second order* in the inverse layer width as $\sim V^2/n^2$ and *not* $\sim V/n$. Indeed, since the quartic coupling v_m in a generic nearly-Gaussian action can take either sign – corresponding to a distribution with either fat tails or thin tails – the leading correction to the entropy must be proportional to minus the *square* of the quartic coupling, $v_m^2\propto V^2/n^2$, in order to guarantee that the entropy decreases.[21] Finally, we see that this correction breaks the perfect additivity for the neurons that we found in the infinite-width limit (A.26). In particular, although the decrease in the entropy is perturbatively small, with an order $1/n^2$ scaling, it also depends *quadratically* on m. This nonlinearity in the number of neurons signals the presence of nontrivial interactions at finite width.

[21] This same argument excludes a contribution of the form $O(u_m)=O\left(U/n^2\right)$ coming from the sextic coupling, cf. (A.55), and similarly excludes any linear contributions from other higher-order couplings.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

Accordingly, we can characterize this statistical dependence by computing mutual information between two nonoverlapping sets of neurons, $\mathcal{M}_1 = \{1, \ldots, m_1\}$ and $\mathcal{M}_2 = \{m_1 + 1, \ldots, m_1 + m_2\}$,

$$\begin{aligned} \mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] &\equiv \mathcal{S}[p(\mathcal{M}_1|x)] + \mathcal{S}[p(\mathcal{M}_2|x)] - \mathcal{S}[p(\mathcal{M}_1, \mathcal{M}_2|x)] \\ &= \frac{m_1 m_2}{8} \left[\frac{V^{(\ell)}}{n_{\ell-1}\left(G^{(\ell)}\right)^2}\right]^2 + O\left(\frac{1}{n^3}\right) , \end{aligned} \tag{A.53}$$

where we used our entropy formula (A.52) in three different ways and also restored layer indices to better interpret this formula. This nonzero mutual information signifies that – at finite width, *only* – for a given layer ℓ, observing the activities of a group of neurons $\mathcal{M}_1$ will convey information about the activities of another group of neurons $\mathcal{M}_2$. This can be thought of as an information-theoretic reformulation and generalization of the *Hebbian learning* principle that we saw for the conditional variance (6.78) in §6.4.1: there we more simply saw that the variance of one neuron $z_2^{(\ell)}$, conditioned on an atypical observation of a second neuron, $z_1^{(\ell)} = \tilde{z}_1^{(\ell)}$, will itself be atypical; here, we can directly characterize how much one group of neurons can know about another nonoverlapping group.[22]

Finally, remembering our *scaling law* for the normalized vertex (5.128), we see that the mutual information (A.53) scales with the depth ℓ of the hidden layer *squared*:

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] \propto \ell^2/n^2. \tag{A.54}$$

In terms of *RG flow*, this means that the mutual information is *relevant*, suggesting that a growing mutual information is helpful when coarse-graining representations: as the fine-grained features are marginalized over, deeper hidden layers will have a growing set of correlations between groups of neurons. In other words, by increasing the size of the nonlinear subadditive term in the entropy, this reduces the number of independently available degrees of freedom in these deeper layers.[23]

NLO Correction: Optimal Aspect Ratio and Tripartite Information

Our leading-order result for the finite-width mutual information, (A.54), naively grows in depth without bounds, suggesting that deeper is always better. Of course, if the depth becomes too large, this naive answer breaks down as the higher-order terms in the perturbation series start to become important. To understand the mutual information at even greater depths, we'll need to compute the *next-to-leading-order* (NLO) correction.

To push our calculations to the next level, we need to ensure that our two variational constraints, (A.35) and (A.36), are satisfied to next-to-leading-order, $O(1/n^2)$.

[22] More generally, it would be interesting to work out the mutual information for multiple inputs in order to understand its data dependence. In that case, we expect it to be mediated by the multi-input four-point vertex and depend on the details of different groupings of four samples from the dataset $\mathcal{D}$.

[23] It would be interesting to try to interpret this in terms of the *optimal brain damage* of [82] or the *lottery ticket hypothesis* of [83].

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

In principle, this means that we will need to also include an $O(1/n^2)$ *sextic* term in our variational action $S_{\rm I}$ as

$$S_{\rm I}(z_1,\ldots,z_m) = \frac{1}{2}\left(g_m - \frac{1}{G}\right)\sum_{i=1}^{m} z_i^2 - \frac{1}{8}v_m \sum_{i,j=1}^{m} z_i^2 z_j^2 + \frac{1}{24}u_m \sum_{i,j,k=1}^{m} z_i^2 z_j^2 z_k^2 + O\left(\frac{1}{n^3}\right). \tag{A.55}$$

Such a sextic term was originally introduced in (ε.18); here, we've specialized it to focus on m preactivations in a layer ℓ, with u_m the running sextic coupling.

Now, let's explain how to explicitly satisfy the variational constraints. First, just as we did before for the entropy in (A.45), note that for a general observable we can express its full expectation in terms of simpler Gaussian expectations as

$$\begin{aligned}\mathbb{E}\left[\mathcal{O}\right] &\equiv \frac{1}{Z}\int\left[\prod_{i=1}^{m} dz_i\right] e^{-S}\mathcal{O} = \left(\frac{Z}{Z_G}\right)^{-1}\left\langle\!\!\left\langle e^{-S_{\rm I}}\mathcal{O}\right\rangle\!\!\right\rangle_G \\ &= \left\langle\!\!\left\langle\mathcal{O}\right\rangle\!\!\right\rangle_G - \left[\left\langle\!\!\left\langle\mathcal{O}S_{\rm I}\right\rangle\!\!\right\rangle_G - \left\langle\!\!\left\langle\mathcal{O}\right\rangle\!\!\right\rangle_G\left\langle\!\!\left\langle S_{\rm I}\right\rangle\!\!\right\rangle_G\right] \\ &\quad + \frac{1}{2}\left[\left\langle\!\!\left\langle\mathcal{O}S_{\rm I}^2\right\rangle\!\!\right\rangle_G - 2\left\langle\!\!\left\langle\mathcal{O}S_{\rm I}\right\rangle\!\!\right\rangle_G\left\langle\!\!\left\langle S_{\rm I}\right\rangle\!\!\right\rangle_G - \left\langle\!\!\left\langle\mathcal{O}\right\rangle\!\!\right\rangle_G\left\langle\!\!\left\langle S_{\rm I}^2\right\rangle\!\!\right\rangle_G + 2\left\langle\!\!\left\langle\mathcal{O}\right\rangle\!\!\right\rangle_G\left\langle\!\!\left\langle S_{\rm I}\right\rangle\!\!\right\rangle_G^2\right] \\ &\quad + O\left(\frac{1}{n^3}\right),\end{aligned} \tag{A.56}$$

where in the last equality, we expanded $\left\langle\!\!\left\langle e^{-S_{\rm I}}\mathcal{O}\right\rangle\!\!\right\rangle_G$ in $S_{\rm I}$ and also used the formula that we already evaluated in (A.43) for the ratio of partition functions. Physically, the terms within the first square brackets say that in an interacting theory, the leading variational correction to an observable is given by the correlation of the observable with the variational part of the action.

To satisfy our first constraint for the metric, (A.35), we can use this general expression (A.56) with the observable

$$\mathcal{O} \equiv \frac{1}{m}\sum_{i=1}^{m} z_i^2, \tag{A.57}$$

for which we can use our constraint (A.35) to easily see that $\mathbb{E}\left[\mathcal{O}\right] = G$. Evaluating this same expectation using the variational sextic action, (A.55), we find

$$\begin{aligned}G &= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left[z_i^2\right] \\ &= G - \left(g_m - \frac{1}{G}\right)G^2 + \frac{(m+2)}{2}v_m G^3 + \left(g_m - \frac{1}{G}\right)^2 G^3 - \frac{(m+2)(m+4)}{4}u_m G^4 \\ &\quad - \frac{3(m+2)}{2}\left(g_m - \frac{1}{G}\right)v_m G^4 + \frac{(m+2)(m+3)}{2}v_m^2 G^5 + O\left(\frac{1}{n^3}\right).\end{aligned} \tag{A.58}$$

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

In evaluating this, you might again find the formula (A.50) helpful. This expression, (A.58), shows how the interacting theory modifies the two-point correlator. Rearranging terms to solve for g_m order by order, we find an NLO version of (A.47), which determines the variational quadratic coupling in terms of the metric and the other higher-order couplings:

$$g_m = \frac{1}{G} + \frac{(m+2)}{2} v_m G - \frac{(m+2)(m+4)}{4} u_m G^2 + \frac{(m+2)}{2} v_m^2 G^3 + O\left(\frac{1}{n^3}\right). \quad \text{(A.59)}$$

Next, to satisfy the second constraint for the four-point vertex, (A.36), we can use our general expression (A.56) with the observable

$$\mathcal{O} \equiv \frac{1}{m(m+2)} \sum_{i,j=1}^{m} z_i^2 z_j^2, \quad \text{(A.60)}$$

for which we can use both our constraints (A.35) and (A.36) to show that $\mathbb{E}[\mathcal{O}] = G^2 + V/n$. Now evaluating $\mathcal{O}$ according to the variational sextic action, (A.55), we get

$$G^2 + \frac{V}{n} = \frac{1}{m(m+2)} \sum_{i,j=1}^{m} \mathbb{E}\left[z_i^2 z_j^2\right] \quad \text{(A.61)}$$

$$\begin{aligned} &= G^2 - 2\left(g_m - \frac{1}{G}\right) G^3 + 3\left(g_m - \frac{1}{G}\right)^2 G^4 + (m+3) v_m G^4 - \frac{(m+4)^2}{2} u_m G^5 \\ &\quad - 4(m+3)\left(g_m - \frac{1}{G}\right) v_m G^5 + \frac{(5m^2+34m+60)}{4} v_m^2 G^6 + O\left(\frac{1}{n^3}\right) \\ &= G^2 + v_m G^4 - (m+4) u_m G^5 + \frac{(m+8)}{2} v_m^2 G^6 + O\left(\frac{1}{n^3}\right), \end{aligned} \quad \text{(A.62)}$$

where to get to the second line we again made heavy use of our formula (A.50), and to get to the final line we plugged in (A.59) for the quadratic coupling. Rearranging terms to solve for v_m order by order, we find an NLO version of (A.48), which determines the variational quartic coupling in terms of the two constraints and the sextic coupling:

$$v_m = \frac{V}{nG^4} - \frac{(m+8)}{2}\left(\frac{V}{nG^3}\right)^2 + (m+4) u_m G + O\left(\frac{1}{n^3}\right). \quad \text{(A.63)}$$

We now finally have all the pieces we need to evaluate the NLO correction to the entropy (A.45).[24] First, let's reevaluate the variance of the variational action to NLO by repeated use of the formula (A.50):

[24] In principle, at this order we would need to continue and find an expression for u_m in terms of an additional six-point vertex constraint, but as we will soon see, u_m doesn't factor into any of our expressions for the mutual information. This is the reason why we are able to analyze these next-to-leading-order corrections without otherwise evaluating additional MLP recursions.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

$$\begin{aligned}
&\left\langle\!\left\langle S_{\mathrm{I}}^2 \right\rangle\!\right\rangle_G - \left\langle\!\left\langle S_{\mathrm{I}} \right\rangle\!\right\rangle_G^2 \qquad \text{(A.64)}\\
&\quad = \frac{m}{2}\left(g_m - \frac{1}{G}\right)^2 G^2 - \frac{m(m+2)}{2}\left(g_m - \frac{1}{G}\right) v_m G^3 + \frac{m(m+2)(m+3)}{8} v_m^2 G^4 \\
&\qquad + \frac{m(m+2)(m+4)}{4}\left(g_m - \frac{1}{G}\right) u_m G^4 - \frac{m(m+2)(m+4)^2}{8} v_m u_m G^5 + O\left(\frac{1}{n^4}\right) \\
&\quad = \frac{m(m+2)}{8}\left[\left(v_m G^2\right)^2 - 2(m+4)\left(v_m G^2\right)\left(u_m G^3\right)\right] + O\left(\frac{1}{n^4}\right) \\
&\quad = \frac{m(m+2)}{8}\left[\left(\frac{V}{nG^2}\right)^2 - (m+8)\left(\frac{V}{nG^2}\right)^3\right] + O\left(\frac{1}{n^4}\right).
\end{aligned}$$

Here, to get to the penultimate line, we used our NLO expression for the quadratic coupling (A.59), and then to get to the final line, we used our NLO expression for the quartic coupling (A.63). Second, let's similarly evaluate the subleading term in our expression (A.45) for the entropy:

$$\begin{aligned}
&\left\langle\!\left\langle S_{\mathrm{I}}^3 \right\rangle\!\right\rangle_G - 3\left\langle\!\left\langle S_{\mathrm{I}}^2 \right\rangle\!\right\rangle_G \left\langle\!\left\langle S_{\mathrm{I}} \right\rangle\!\right\rangle_G + 2\left\langle\!\left\langle S_{\mathrm{I}} \right\rangle\!\right\rangle_G^3 \qquad \text{(A.65)}\\
&= m\left(g_m - \frac{1}{G}\right)^3 G^3 - \frac{9m(m+2)}{4}\left(g_m - \frac{1}{G}\right)^2 v_m G^4 \\
&\quad + \frac{3m(m+2)(m+3)}{2}\left(g_m - \frac{1}{G}\right) v_m^2 G^5 - \frac{m(m+2)(5m^2+34m+60)}{16} v_m^3 G^6 + O\left(\frac{1}{n^4}\right) \\
&= -\frac{m(m+2)(m+8)}{8}\left(v_m G^2\right)^3 + O\left(\frac{1}{n^4}\right) \\
&= -\frac{m(m+2)(m+8)}{8}\left(\frac{V}{nG^2}\right)^3 + O\left(\frac{1}{n^4}\right).
\end{aligned}$$

Putting (A.64) and (A.65) back into our variational expression for the entropy (A.45), we finally arrive at

$$\begin{aligned}
&\mathcal{S}[p(z_1, \ldots, z_m|x)] \qquad \text{(A.66)}\\
&\quad = \frac{m}{2}\log(2\pi e G) - \frac{(m^2+2m)}{16}\left(\frac{V}{nG^2}\right)^2 + \frac{(m^3+10m^2+16m)}{48}\left(\frac{V}{nG^2}\right)^3 + O\left(\frac{1}{n^4}\right).
\end{aligned}$$

This result in turn lets us compute the NLO correction to the mutual information between two sets of neurons, $\mathcal{M}_1 = \{1, \ldots, m_1\}$ and $\mathcal{M}_2 = \{m_1 + 1, \ldots, m_1 + m_2\}$:

$$\begin{aligned}
&\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] \qquad \text{(A.67)}\\
&\quad \equiv \mathcal{S}[p(\mathcal{M}_1|x)] + \mathcal{S}[p(\mathcal{M}_2|x)] - \mathcal{S}[p(\mathcal{M}_1, \mathcal{M}_2|x)] \\
&\quad = \frac{m_1 m_2}{8}\left[\frac{V^{(\ell)}}{n_{\ell-1}\left(G^{(\ell)}\right)^2}\right]^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48}\left[\frac{V^{(\ell)}}{n_{\ell-1}\left(G^{(\ell)}\right)^2}\right]^3 + O\left(\frac{1}{n^4}\right),
\end{aligned}$$

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

where once again we have restored layer indices on the metric, the four-point vertex, and the layer width. Note that as promised, the sextic coupling u_m dropped out in the end.[25]

Excitingly, these two terms have opposite signs. To explain our excitement, let us plug in our scaling solution for the normalized four-point vertex (5.128) evaluated at the final layer $\ell = L$:

$$\frac{V^{(L)}}{n_{L-1}\left(G^{(L)}\right)^2} \equiv \nu r. \tag{A.68}$$

Here, $r \equiv L/n$ is the overall depth-to-width aspect ratio of the network, and ν is an activation-function-dependent constant: for the $K^\star = 0$ universality, we have (5.131)

$$\nu = \frac{2}{3}, \tag{A.69}$$

independent of the details of the activation function itself; for scale-invariant activation functions, we have (5.130)

$$\nu = \left(\frac{3A_4}{A_2^2} - 1\right), \tag{A.70}$$

with $A_2 \equiv (a_+^2 + a_-^2)/2$ and $A_4 \equiv (a_+^4 + a_-^4)/2$.[26] Plugging this scaling solution (A.68) into our NLO expression for the mutual informtion (A.67), we get

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] = \frac{m_1 m_2}{8}\nu^2 r^2 - \frac{m_1 m_2(20 + 3m_1 + 3m_2)}{48}\nu^3 r^3 + O\left(r^4\right). \tag{A.71}$$

Now, let us explain our excitement. At one extreme, in the infinite-width limit $r \to 0$, this mutual information vanishes, as we already knew from (A.27). Thus, for small aspect ratios $r \ll 1$, the first leading-order term dominates, and the mutual information increases as depth increases. As the depth continues to increase, then the second term begins to dominate, *decreasing* the mutual information. But we know by the *subadditivity* of the entropy, (A.15), that the mutual information is always bounded from below by zero, and so for large enough aspect ratios r, this decreasing will be balanced by additional higher-order corrections. Taken altogether, we expect that at some nonzero but not too large r, the mutual information will reach a local maximum. Indeed, maximizing (A.71), we find an **optimal aspect ratio** for the network:

[25] This is another realization of the principle of maximum entropy for nearly-Gaussian distributions. In particular, the entropy for a zero-mean distribution that satisfies constraints fixing its two-point correlator and its connected four-point correlator – and otherwise leaves unfixed its connected six-point correlator – is given to order $1/n^3$ by our expression (A.66). Thus, if we did add a third constraint that fixed the connected six-point correlator, with $U = O\left(1/n^2\right)$, then any terms of the form $O(v_m u_m) = O\left(UV/n^3\right)$ cannot appear in (A.66), as they could *increase* the entropy depending on the sign of UV.

[26] This gives $\nu = 2$ for `linear` activations and $\nu = 5$ for the `ReLU`.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

$$r^\star = \left(\frac{4}{20 + 3m_1 + 3m_2}\right)\frac{1}{\nu}, \tag{A.72}$$

with ν containing all of the details of the activation function.[27]

Although it's not immediately obvious, maximizing a mutual information such as (A.68) is closely related to well-known **unsupervised learning** objectives.[28] In contrast to a *supervised* learning setting where the goal is to predict the true output $y \equiv f(x)$ for any input sample x, in an *unsupervised* learning setting, the goal is to learn representations for a collection of input samples by observing patterns in the data. For human-generated datasets, this has the advantage of eliminating the tedious task of labeling all the samples. It should also be no surprise, given the benefit of representation learning (§11), that models (pre)trained with unsupervised learning algorithms can often be efficiently fine-tuned on subsequent supervised learning tasks.

In the current context, rather than doing any actual learning, we are understanding, a priori, which choice of architecture and activation function can lead to a larger mutual information between neurons in deeper layers.[29] In particular, comparing the values of ν in (A.69) and (A.70) for different choices of activation function, we see in principle that networks built from the `tanh` activation function should be deeper than networks built from `ReLU` activation functions to have the same mutual information in a layer.

With this objective in mind, we should continue maximizing (A.71) across all the possible partitions (m_1, m_2). This picks out a very natural choice of a partition spanning the whole layer and of equal sizes: $m_1 = m_2 = n_L/2$. With this further maximization, we get

$$r^\star = \left(\frac{4}{20 + 3n_L}\right)\frac{1}{\nu} \tag{A.73}$$

for the optimal aspect ratio and

$$\mathcal{I}\left[p(\mathcal{M}_1, \mathcal{M}_2|x)\right] = \frac{1}{6}\left(\frac{n_L}{20 + 3n_L}\right)^2 \tag{A.74}$$

for the associated value of the maximized mutual information. In particular, this corresponds to a lower bound on the *InfoMax* objective that we discussed in footnote 28.[30]

[27] Don't worry about the higher-order contributions $O\left(r^4\right)$ that we neglected in obtaining the solution (A.72): for a particular choice of activation function, i.e., for a choice of ν, the estimate of the optimal value (A.72) is *a posteriori* justified so long as the product $\nu r^\star$ is perturbatively small for a particular ν and a particular grouping of neurons (m_1, m_2). FYI, this argument is analogous the one given to justify the two-loop *Banks–Zaks fixed point* of the renormalization group in non-Abelian gauge theory [84].

[28] In particular, the *InfoMax principle* [85] recommends maximizing the mutual information between the input x and a representation $z(x)$. A related notion involves maximizing the mutual information between different representations, $z_1(x)$ and $z_2(x)$, for the same input x [86]. This latter notion can be shown to lower bound the InfoMax objective and thus motivates our analysis here.

[29] Similarly, we may think of our criticality analysis as a type of unsupervised learning or pretraining criterion in which we are understanding, a priori, which choice of initialization hyperparameters leads to an ensemble that generalizes most robustly after training, cf. §10.3.

[30] While this value (A.74) may seem small, remember that our analysis is based entirely on the prior distribution before any learning has taken place.

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press

As a final comment on (A.73), we can think of this optimal aspect ratio as defining a scale that separates the effectively-deep regime for which our effective theory is valid from the overly-deep regime where the theory is strongly coupled and networks are no longer trainable. We will push this interpretation further in the next appendix when we discuss residual networks.

For a final computation, let's look at the *tripartite information* (A.24) for mutually-exclusive subsystems $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$ of sizes (m_1, m_2, m_3) neurons, respectively. Plugging in (A.66) for various different combinations of the sizes, we find

$$\mathcal{I}_3\left[p(\mathcal{M}_1,\mathcal{M}_2,\mathcal{M}_3|x)\right] = \frac{m_1 m_2 m_3}{8}\left[\frac{V^{(\ell)}}{n_{\ell-1}\left(G^{(\ell)}\right)^2}\right]^3 + O\left(\frac{1}{n^4}\right), \tag{A.75}$$

which, as per (5.128), scales cubically with depth, $\mathcal{I}_3\left[p(\mathcal{M}_1,\mathcal{M}_2,\mathcal{M}_3|x)\right] \propto \ell^3/n_{\ell-1}^3$, and thus indicates that a nonzero result only first appears at $O(1/n^3)$.[31] Importantly, we see that the tripartite information is exclusively *positive*, meaning that any three groups of neurons in a layer will form *redundant* representations under RG flow: knowing the activities of any one group of neurons $\mathcal{M}_1$ means you would learn less information about a second group of neurons $\mathcal{M}_2$ from observing a third group $\mathcal{M}_3$ than you otherwise would have learned had you not already known $\mathcal{M}_1$. It would be interesting to try to understand further how this property relates to the coarse-graining mechanism of the representation group flow to the deeper layers.[32]

[31] In hindsight, it's obvious that we needed the entropy to have at least a cubic dependence on the sizes m_i to find a nonzero answer for the tripartite information, just as we needed the entropy to have at least a quadratic dependence on the sizes to find a nonzero answer for the mutual information (A.53).

[32] Again, it would also be interesting to try and interpret this in terms of the *optimal brain damage* of [82] or the *lottery ticket hypothesis* of [83].

https://doi.org/10.1017/9781009023405.016 Published online by Cambridge University Press