DLT Lecture 2 : Thermodynamics of Singular Learning Theory

In the first lecture it was argued that deep learning is an emerging general purpose technology, and that theory has an important role to play. An analogy was made to the role of thermodynamic theory (or more precisely its precursors) in enabling the development of steam power in the first industrial revolution. This analogy was not idly chosen: thermodynamics and statistical mechanics also provide a useful framework for understanding Watanabe's singular learning theory in general and deep learning theory in particular.

We will assume familiarity with thermodynamics at the level of [Callen] and the basics of singular learning theory [W], and use the language of the former to give an accessible presentation of the latter.

It may seem strange to use thermodynamics (which is expressed in the language of physical systems and concepts like energy, entropy, volume, pressure, etc.) to present a topic in statistics. Surely there isn't anything analogous to pressure in deep learning theory! However it has long been understood that the fundamental principles of thermodynamics extend far beyond gases in bottles and ferromagnets. The justification for this lies outside the scope of this lecture, but you can consider the information-theoretic interpretation of entropy [Callen, §17.1] and the origin of macroscopic thermodynamic coordinates in symmetry-breaking explained in [Callen, §21].

This leads to a precise thermodynamic dictionary for singular learning theory, exhibiting which is the purpose of this lecture.

1. The objects of singular learning theory

For concreteness we consider a triple $(p(y|x,w),\ q(y|x),\ \varphi(w))$ associated to a class of functions $f(x,w)$ as in [MDLG1, §3, Appendix A], so $f: \mathbb{R}^L \times W \to \mathbb{R}^M$ and

$$p(y|x,w) = \frac{1}{(2\pi)^{M/2}} \exp\left(-\tfrac{1}{2}\| y - f(x,w)\|^2\right). \tag{2.1}$$

Here $q(y|x) = p(y|x,w_0)$ for some $w_0 \in W$ (i.e. we assume the true distribution is realisable) and $\varphi(w)$ is a prior distribution on a compact space $W \subseteq \mathbb{R}^d$ of weights. The case where $f(x,w)$ is a neural network is of interest for deep learning theory, but nothing we will say is specific to this case, and the dictionary we develop holds for a general class of models, not just those of exponential form (2.1).

The central function in Singular Learning Theory (SLT) is the Kullback–Leibler (KL) divergence ("distance") between the true distribution and the model

$$K(w) = \int\int q(y|x) \log \frac{q(y|x)}{p(y|x,w)} q(x)\, dx dy. \tag{2.2}$$

In the situation of (2.1) we compute (see SLT8 p.3)

$$K(w) = \frac{1}{2} \int \| f(x,w) - f(x,w_0)\|^2 q(x)\, dx \tag{2.3}$$

which is clearly a measure of the <u>error</u> associated to $w \in W$. In practice we can only interact with the true distribution by drawing samples $D_n = \{(x_i, y_i)\}_{i=1}^n$ and the <u>empirical KL divergence</u> associated to such a sample is [W, p. 5]

$$K_n(w) = \frac{1}{n} \sum_{i=1}^n \log \frac{q(y_i|x_i)}{p(y_i|x_i)} \tag{2.4}$$

$$= -S_n + L_n(w)$$

where $S_n = -\frac{1}{n}\sum_{i=1}^{n} \log q(y_i|x_i)$ is the empirical entropy and

$$L_n(w) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i,w) \tag{3.1}$$

is the "log loss" or negative log likelihood (we follow the notation of [WAIC] here as [W] uses $L_n(w)$ to denote the likelihood. Note that $e^{-nL_n(w)} = \prod_{i=1}^{n} p(y_i|x_i,w)$ is the likelihood, modulo factor of $q(x_i)$). In practice we do not have access to q and we may not be able to calculate K_n or S_n, but we can compute $L_n(w)$ which is maximised when $K_n(w)$ is minimised (as q is fixed). In the situation of (2.1)

$$\begin{aligned} L_n(w) &= -\frac{1}{n}\sum_{i=1}^{n} \log\left[\frac{1}{(2\pi)^{M/2}} \exp\left(-\tfrac{1}{2}\|y_i - f(x_i,w)\|^2\right)\right] \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left[-\frac{M}{2}\log(2\pi) - \tfrac{1}{2}\|y_i - f(x_i,w)\|^2\right] \qquad (3.2) \\ &= \frac{M}{2}\log(2\pi) + \frac{1}{n}\sum_{i=1}^{n} \tfrac{1}{2}\|y_i - f(x_i,w)\|^2 \end{aligned}$$

is up to a constant that depends only on M, and the factor $1/2$, the <u>mean-squared error</u>. If $\mathbb{E}[K(w)] < \infty$ then we have convergence in probability $K_n(w) \to K(w)$ for all $w \in W$, and if $\mathbb{E}[K(w)^2] < \infty$ then $L_n(w) \to K(w) - S$ in probability where S is the entropy of the true distribution [W, p. 6]. So heuristically

$$K_n = \text{const} + \tfrac{1}{2}\,\text{MSE} - S_n \tag{3.3}$$

The next important object is the <u>posterior</u>, which using Bayes rule is

$$p(w|D_n) = \frac{p(D_n|w)\,p(w)}{p(D_n)} \tag{3.4}$$

We have decided to write $\varphi(w)$ for $p(w)$.

Hence

$$p(w|D_n) = \frac{1}{p(D_n)}\varphi(w)\prod_{i=1}^{n} p(x_i,y_i|w)$$
$$= \frac{1}{p(D_n)}\varphi(w)\prod_{i=1}^{n} p(y_i|x_i,w)\,q(x_i) \qquad (4.1)$$

Notice that we model only the conditional distribution $p(y|x,w)$, so we assume $q(x)$ is given and set $p(x,y|w) = p(y|x)\,q(x)$. We note that the factor $\prod_{i=1}^{n} q(x_i)$ does not depend on w, so

$$p(w|D_n) = \frac{\prod_{i=1}^{n} q(x_i)}{p(D_n)}\varphi(w)\,e^{-nL_n(w)}$$
$$= \frac{1}{Z_n}\varphi(w)\,e^{-nL_n(w)} \qquad (4.2)$$

where using that $p(w|D_n)$ is a probability distribution

$$Z_n = \int dw\,\varphi(w)\,e^{-nL_n(w)}. \qquad (4.3)$$

This quantity is called the evidence. Using (2.4) we can arrive at an equivalent formulation of the posterior which is closer to thermodynamics. By (4.2),

$$p(w|D_n) = \frac{1}{Z_n}\varphi(w)\,e^{-nK_n(w)}\,e^{-nS_n}$$
$$= \frac{1}{Z_n^0}\varphi(w)\,e^{-nK_n(w)} \qquad (4.4)$$

where $Z_n^0 = e^{nS_n}Z_n = \int dw\,\varphi(w)\,e^{-nK_n(w)}$ is called the normalised evidence [W, p. 20]. So in summary: given a dataset D_n, $K_n(w)$ is an empirical estimate of how close $p(y|x,w)$ is to $q(y|x)$, and the posterior $p(w|D_n)$ puts more probability density near w the smaller this estimate is (ignoring φ).

Remark As a KL divergence $K(w)$ is always non-negative and by definition the quantity $L_n(w)$ is non-negative, but it is possible $K_n(w) < 0$ (although $K_n(w) \geq -S_n$). For our purposes the difference between using L_n or K_n is just a shift in the energy by a constant S_n independent of w, and either will do (but note S_n is a random variable as it depends on D_n).

2. The objects of thermodynamics

Thermodynamics is the study of macroscopic observables (such as energy, volume, pressure and mole numbers) associated to equilibrium states, and its central problem is to predict the equilibrium state which eventually results after a constraint is removed from a system currently in equilibrium [Callen, §1]. The principle governing the theory is that there is a quantity, the entropy S, which is a function of the energy U, volume V, and mole number(s) N, and which is maximised at equilibrium. The fundamental relation is this functional dependence

$$S = S(U, V, N) \tag{5.1}$$

Every conceivable thermodynamic attribute of the system is known once this fundamental relation is known [Callen, p.28]. Since $\partial S/\partial U > 0$ by hypothesis, it is equivalent by the inverse function theorem to know the fundamental relation in the following form

$$U = U(S, V, N) \tag{5.2}$$

The quantities U, S, V, N are called extensive since they are additive over component subsystems. The intensive thermodynamic parameters do not have this property, and are defined as partial derivatives of either (5.1) or (5.2).

$$T = \frac{\partial U}{\partial S} \qquad \text{temperature}$$
$$P = -\frac{\partial U}{\partial V} \qquad \text{pressure} \qquad (6.1)$$
$$\mu = \frac{\partial U}{\partial N} \qquad \text{electrochemical potential}$$

$$dU = TdS - PdV + \mu dN$$

To see that these "definitions" of T, P agree with your background knowledge of temperature and pressure see [Callen, §2-4, §2-5, §2-7].

Example In a simple ideal gas [Callen, 3-4]

$$S = Ns_0 + NR\ln\left(\left[\frac{U}{U_0}\right]^c\left[\frac{V}{V_0}\right]\left[\frac{N}{N_0}\right]^{-(c+1)}\right) \qquad (6.2)$$

where c is a constant depending on the gas (although it is constant across "similar" gases) and R is the "universal" gas constant.

In practice thermodynamics is often done in one or another "representations" corresponding to Legendre transformations of U. Mathematically speaking this is only valid locally (since e.g. U need not be globally convex w.r.t. V) and this subtlety is (in my opinion) the main reason thermodynamics texts are hard for mathematicians to read, so be careful! [Callen, §5]. Nonetheless this method is remarkably powerful in treating systems in contact with heat or pressure reservoirs. It will turn out that the Gibbs representation is the one most appropriate to SLT.

It is far from obvious what the appropriate extensive thermodynamic parameters are for SLT. So we instead begin with the "microscopic" degrees of freedom and the statistical mechanics of SLT [Callen, III] from which U, S, V, N emerge as averages.

3. Statistical mechanics of singular learning theory

We formulate SLT as a "Gibbs ensemble" [A, §4.6.5], [PLOW] in which we imagine the learning process as a physical system in contact with a thermal and pressure reservoir. At the moment it is not meant to be clear what this means; this we leave to later. The physical system consists of a space of states and the Hamiltonian which assigns to each state its energy.

Def^n The space of (microscopic) states is W, viewed as a measure space with the Lebesgue measure dw

It is standard in the statistical mechanical point of view on optimisation theory to take the Hamiltonian to be the loss, which in our case is $nL_n(w)$ or $nK_n(w)$ (cf. (3.2)). This "random" Hamiltonian depends on the random variable D_n and measures the violation of the primary constraint (modelling the true distribution) but we should not ignore the prior $\varphi(w)$ which effectively imposes a secondary constraint (if $\varphi(w)$ vanishes outside some radius, or is just extremely small, we have constrained the search for solutions to states within that radius).

In short, the prior plays the role of walls containing a gas (at least insofar as the walls constrain the possible values of position coordinates of gas particles).

Example Suppose $W = \mathbb{R}^d$ (ignore that this is not compact, it's not important right now), and

$$\begin{aligned} \varphi(w) &= \frac{1}{(2\pi\beta)^{d/2}} \exp\left(-\frac{1}{2}\frac{\|w\|^2}{\beta^2}\right) \\ -\log\varphi(w) &= \frac{d}{2}\log(2\pi\beta) + \frac{1}{2}\cdot\frac{1}{\beta^2}\cdot\sum_{j=1}^{d} w_j^2 \end{aligned} \tag{7.1}$$

The negative log probability acts like a potential energy with associated force $-\frac{\partial}{\partial w_j}(-\log\varphi(w)) = -\frac{1}{\beta^2}w_j$, i.e. a simple harmonic oscillator with spring constant β^{-2}.

The (random) Hamiltonian incorporating both energies is

$$H_n(w) = nK_n(w) - \frac{1}{\beta}\log \mathcal{G}(w). \tag{8.1}$$

In the situation of (2.1) and (7.1) this is

$$H_n(w) = \text{const.} + \sum_{i=1}^{n} \tfrac{1}{2}\| y_i - f(x_i, w)\|^2 + \tfrac{1}{\beta}\cdot\tfrac{1}{2}\cdot\tfrac{1}{6^2}\cdot\sum_{j=1}^{d} w_j^2$$

According then to the standard derivation [A, §4.6.5] of the Gibbs ensemble under some constraints which dictate that the state lies in $W' \subseteq W$ the probability density associated to the system being in state w is

$$p^{\text{Boltz}}(w \mid D_n)\,dw = \frac{1}{Z^{\text{Boltz}}} e^{-\beta H_n(w)}\,dw \tag{8.2}$$

where $Z^{\text{Boltz}} = \int_{W'} e^{-\beta H_n(w)}\,dw$ is the normalisation constant. This is called the Boltzmann distribution, where $\beta = 1/T$ is the inverse temperature. Clearly when $\beta = 1$ and $W' = W$ we recover the Bayesian posterior (4.4)

$$\begin{gathered} p^{\text{Boltz}}(w \mid D_n)\,dw = p(w \mid D_n)\,dw \\ Z^{\text{Boltz}} = Z_n^{\circ} \end{gathered} \tag{8.3}$$

In physics the partition function Z^{Boltz} is viewed as a function of T, V, N by setting $\beta = 1/T$ and restricting the integral to states $W' \subseteq W$ consistent with the given values V, N of the other thermodynamic parameters. The dependence on T, N poses no problem from the statistical point of view but the meaning of V is currently unclear.

Inspired by the role of temperature in the Boltzmann distribution we introduce the tempered posterior [W, p. 20]

$$p^\beta(w|D_n) = \frac{1}{Z_n^\circ}\varphi(w)\, e^{-n\beta K_n(w)} \tag{9.1}$$

$$Z_n^\circ = Z_n^\circ(\beta) = \int dw\, \varphi(w)\, e^{-n\beta K_n(w)}$$

so that (8.3) holds as an equality for any β between the Boltzmann distribution and the tempered posterior (note that this agreement motivates the $1/\beta$ in (8.1)).

Def^n We denote the expectation of $f(w)$ with respect to the tempered posterior by

$$\mathbb{E}_w^\beta[f(w)] = \frac{1}{Z_n^\circ}\int_W f(w)\varphi(w)\, e^{-n\beta K_n(w)}\, dw \tag{9.2}$$

From a physical point of view the "learning machine" at equilibrium has a macroscopic state characterised by extensive parameters to be introduced in the next section, but at a microscopic level undergoes rapid transitions between states of many different (microscopic, i.e. per microstate) energies given by H_n. This picture is not as abstract as it may seem: this system is very similar to the in silico system which attempts to generate samples from the posterior using Hamiltonian Monte Carlo.

4. Average energy

By taking averages with respect to the Boltzmann distribution we can finally introduce the macroscopic thermodynamic parameters implicit in SLT.

The average energy U of the system introduced above is

$$\begin{aligned} U &= \int dw\; p^{Boltz}(w \mid D_n)\; H_n(w) \\ &= \int dw\; p^{\beta}(w \mid D_n)\Big[nK_n(w) - \tfrac{1}{\beta} \log \varphi(w) \Big] \qquad (10.1) \\ &= n\, \mathbb{E}_w^{\beta}\big[K_n(w) \big] + 1/\beta\; \mathbb{E}_w^{\beta}\big[-\log \varphi(w) \big] \end{aligned}$$

The quantity $G_t(n, \beta) = \mathbb{E}_w^{\beta}[K_n(w)]$ is what Watanabe calls the <u>Gibbs training error</u>, which in light of (3.2), (3.3) is in the case of an exponential class of models (2.1) the contribution to the average energy U from the loss and the empirical entropy [W, Def^n 1.8]. [W, § 6.3.1].

The quantity $\mathbb{E}_w^{\beta}[-\log \varphi(w)]$ is the contribution to the energy from interaction with the prior — this term will be large if for example $\varphi(w)$ is Gaussian (see (7.1)) with a small variance, but our particles (think Markov chains) spend a lot of time far away from the origin where the "force" exerted by the prior is large.

We now examine both contributions to U in more detail.

4.1 Gibbs training error

By (2.4), $K_n(w) = -S_n + L_n(w)$ so by [WAIC, Theorem 4] we have

$$n\mathbb{E}_w^\beta[K_n(w)] = nK_n(w_0) + \frac{\lambda}{\beta} + U_n\sqrt{\lambda/2\beta} + O_p(1) \qquad (10.5.1)$$

where U_n is a sequence of random variables, $\mathbb{E}[U_n] = 0$ and $\mathbb{E}[(U_n)^2] < 1$ (here $\mathbb{E}[-]$ means expectation with respect to the dataset) and w_0 is a true parameter.

The indicated term gives a contribution λT to the energy U.

Note that this quantity is still a random variable depending on the dataset D_n.

To put this in a more familiar form, recall [W, Theorem 6.8] that there is a random variable G_t^* such that $nG_t^\beta(n) \longrightarrow G_t^*$ converges in law as $n \to \infty$, and by [W, Theorem 6.10] (here expectations are with respect to D_n)

$$\mathbb{E}[G_t^*] = \frac{\lambda}{\beta} - \nu(\beta)$$

where λ is the RLCT and $\nu(\beta)$ is the singular fluctuation.

Problem 1 What is the physical meaning of $\nu(\beta)$?

4.2 Configuration entropy of W_0

Let us examine the $\mathbb{E}_w^\beta[\log \mathcal{G}(w)]$ term in the case where $(p, q, \mathcal{G})$ is regular. By definition

$$\mathbb{E}_w^\beta[\log \mathcal{G}(w)] = \frac{1}{Z_n^\circ}\int \log(\mathcal{G}(w))\,\mathcal{G}(w)\, e^{-n\beta K_n(w)}\, dw$$

We can write [W, Remark 1.14] in regular case $K_n(u_\alpha) = K(u_\alpha) - \frac{\xi_n(u_\alpha)}{\sqrt{n}}\cdot u_\alpha$ in a local coordinate $u_\alpha = (u_{\alpha,1}, \ldots, u_{\alpha,d})$ around the points w_α of $W_0 = \{w_\alpha\}_\alpha$. Then as $n \to \infty$ we have by the Laplace approximation

$$= \frac{1}{Z_n^\circ}\sum_\alpha \int_{U_\alpha} \log(\mathcal{G}(u_\alpha))\,\mathcal{G}(u_\alpha)\, e^{-n\beta K(u_\alpha) + \sqrt{n}\beta\,\xi_n(u_\alpha)\cdot u_\alpha}\, du_\alpha$$

where U_α is a small neighbourhood of w_α. Applying the Laplace transformation again, and ignoring ξ since it contributes only a linear term, we obtain

$$\approx \frac{1}{Z_n^\circ}\left(\frac{2\pi}{n}\right)^{d/2}\sum_\alpha \frac{\mathcal{G}(w_\alpha)\log\mathcal{G}(w_\alpha)}{|H(\beta K)(w_\alpha)|^{1/2}}$$

Supposing the eigenvalues of the Hessian near w_α are $\lambda_\alpha^{(1)}, \ldots, \lambda_\alpha^{(d)}$

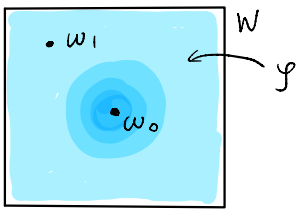
$$= \frac{1}{Z_n^\circ}\left(\frac{2\pi}{n}\right)^{d/2}\beta^{d/2}\sum_\alpha \frac{\mathcal{G}(w_\alpha)\log\mathcal{G}(w_\alpha)}{\sqrt{\lambda_\alpha^{(1)}\cdots\lambda_\alpha^{(d)}}} \qquad (11.1)$$

Assuming for simplicity that $\lambda_\alpha^{(i)} = \lambda^{(i)}$ is independent of α for $1 \le i \le d$ and setting $A = \sum_\alpha \mathcal{G}(w_\alpha)$ we have

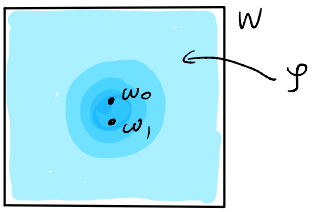
$$\mathbb{E}_w^\beta[\log \varphi(w)] \approx \frac{\beta^{d/2}}{Z_n^\circ(\beta)}\left(\frac{2\pi}{n}\right)^{d/2}\frac{A}{\sqrt{\lambda^{(1)}\cdots\lambda^{(d)}}}\left\{\log A - S_\varphi\right\} \tag{12.1}$$

where $\hat{\varphi}(w_\alpha) = \varphi(w_\alpha)/A$ and $S_\varphi = -\sum_\alpha \hat{\varphi}(w_\alpha)\log\hat{\varphi}(w_\alpha)$.

The quantity S_φ can be thought of as a <u>configuration entropy</u> of W_0 against the prior. To see this consider two examples, in which $W_0 = \{w_0, w_1\}$



S_φ small
$\hat{\varphi}(w_1) \ll \hat{\varphi}(w_0)$



S_φ large
$\hat{\varphi}(w_0) \approx \hat{\varphi}(w_1)$

In the singular case W_0 is a complicated analytic variety, but arguably we can still interpret $\mathbb{E}_w^\beta[\log \varphi(w)]$ as a measure of the entropy of how W_0 is configured relative to φ. Note that as n increases, $e^{-nK_n(w)}$ dominates any terms bounded away from zero, and so $\mathbb{E}_w^\beta[\log \varphi(w)]$ converges to the configurational entropy in the above sense of $\{w_1, \ldots, w_m\} \subseteq W_0$ where $w_1, \ldots, w_m$ are the finitely many point where the point RLCT equals λ (i.e. the "most singular" points of W_0).

5. Entropy

The entropy of the Boltzmann distribution is

$$S = -\int dw\, p^\beta(w|D_n)\, \log\big(p^\beta(w|D_n)\big)$$

$$= -\mathbb{E}^\beta_w\big[\log\varphi(w) - n\beta K_n(w) - \log Z^\circ_n\big]$$

$$= \underbrace{\mathbb{E}^\beta_w[-\log\varphi(w)]}_{\text{entropy of configuration of } W_0 \text{ w.r.t. } \varphi} + \underbrace{n\beta\,\mathbb{E}^\beta_w[K_n(w)]}_{\text{entropy of most singular point}} - \mathbb{E}^\beta_w[F^\circ_n] \qquad (13.1)$$

where $F^\circ_n = -\log Z^\circ_n$ is the free energy [W, §1.4.2] which can be written as $\lambda\log n - (m-1)\log\log n + F^R_n(\xi)$. Combining this with (10.5.1) gives

$$= \mathbb{E}^\beta_w[-\log\varphi(w)] + n\beta K_n(w_0) + \lambda + U_n\sqrt{\lambda/2\beta} \qquad (13.2)$$

$$+ \lambda\log n - (m-1)\log\log n + F^R_n(\xi) + O_p(1)$$

where the underlined terms seem the most significant. Note that this should be taken with a grain of salt at present, since the $O_p(1)$ term may in principle cancel with other constant terms. Note $\mathbb{E}[F^R_n]$ converges to a constant [W, Corr 6.1].

Modulo these details, (13.2) is our desired fundamental equation (5.1) for the entropy, and by inspecting it together with the form (8.1) of the macroscopic Hamiltonian we are led to the idea that the thermodynamic coordinates in SLT are n and $T = 1/\beta$. We can interpret $1/n$ as a kind of "sampling temperature" which controls the noise scale in the comparison of K_n to K (see e.g. [W, (1.19)]). Other thermodynamic coordinates (e.g. ε in φ) may also be introduced.

Remark For the moment we treat n as continuous (in order to write expressions like $\frac{\partial}{\partial n}$) and leave it to later to make sense of this.

We can choose to move forward with the fundamental equation in the form $S = S(n, \beta)$ or switch to the free energy $F_n^\circ = -\log Z_n^\circ$, which is more standard given our starting point with a microscopic Hamiltonian. In order to derive the extensive thermodynamic coordinates conjugate to the intensive coordinates n, β we consider the partial derivatives

$$\frac{\partial}{\partial n} F_n^\circ \quad , \quad \frac{\partial}{\partial \beta} F_n^\circ . \tag{14.1}$$

Loosely speaking the rate of increase of F_n° with n is the Bayes generalisation error B_g (at least at temperature $\beta = 1$) see [W, Theorem 1.2]. We can examine the other extensive quantity $\frac{\partial}{\partial \beta} F_n^\circ$ using [W, Main Formula II, p. 34], [W, Main Theorem 6.2, p. 174] according to which

$$F_n^\circ = \lambda \log n - (m-1) \log\log n + F_n^R(\xi) + o_p(1) \tag{14.2}$$

where

$$F^R(\xi) = -\log\left(\int du^* \int_0^\infty dt\, t^{\lambda - 1} e^{-\beta t + \sqrt{t}\beta \xi(u)} \right) \tag{14.3}$$

Ignoring the $o_p(1)$ term, we have

$$\frac{\partial}{\partial \beta} F_n^\circ = \frac{\partial}{\partial \beta} F^R(\xi)$$

Problem 2 Understand this quantity $\frac{\partial}{\partial \beta} F_n^\circ$.

The Thermodynamic Dictionary

Thermodynamics	Singular Learning Theory
Microscopic Hamiltonian	$nK_n(w) - \frac{1}{\beta}\log\varphi(w)$
Boltzmann distribution	Bayesian posterior
Entropy S	RLCT λ + ...
Intensive coordinates	n, β
Extensive coordinates	Generalisation error, ?
First-order phase transitions	} see DLT3
Second-order phase transitions	

6. Examples of thinking thermodynamically

6.1 Training and refrigeration

What does it mean to train a neural network from the point of view of this dictionary? Speaking loosely, running SGD at a fixed learning rate is similar to the "burn in" period of a MCMC sampler, so that running SGD for K steps and then returning the weight is similar to sampling from the Bayesian posterior at a fixed temperature (related to the learning rate). The precise relationship between SGD and the posterior is an open problem, but nonetheless this analogy is worth knowing, and for the sake of illustration we will assume the analogy is precise in the following.

Usually in vanilla SGD (i.e. no momentum etc) the learning rate will be "annealed" (i.e. decreased) over the course of training. This corresponds to slowly decreasing the temperature of the learning machine of Section 3 (here "slowly" means quasi-static, i.e. the sampling process has time in between the decrements in T to return to equilibrium). This concentrates the posterior / Boltzmann distribution near true parameters, so that one is more likely to sample "good" models. Decreasing the temperature of the learning machine requires work and somewhere to transfer the excess heat; that is, neural network training is a refrigerator [Callen, S4-6]. This should be taken literally: even though the temperature and entropy of a learning machine may seem abstract, the heat and work involved in refrigerating them is quite real (see Google's datacenter cooling bills).

6.2 A recipe for research problems

The implications of a fully thermodynamic point of view on deep learning are quite profound, because it suggests we should be more ambitious about designing complicated learning machines (engines!).

Meta Problem Take any machine in a thermodynamics textbook and see what it means as a design for a learning machine.

Some examples:

Problem 3 Give a thermodynamic treatment of transfer-learning / fine-tuning as a coupling of two learning machines (see [HKK, §3.4]).

Problem 4 Give a thermodynamic treatment of meta-learning.

See overleaf for some informal usage of terms that could potentially be given a mathematical treatment.

Mergeable Fragments

Similarly to how a software engineer merges together pre-existing libraries (or systems) with their code in order to build useful programs, an ML engineer merges together code fragments, data fragments, analysis fragments and model fragments on a regular basis in order to build useful ML pipelines. A notable difference between software engineering and ML engineering is that even when the code is fixed for the latter, data is usually volatile for it (e.g. new data arrives on a regular basis) and as such the downstream artifacts need to be produced frequently and efficiently. For example, a new version of a model usually needs to be produced if any part of its input data has changed. As such, it is important for ML pipelines to produce artifacts that are mergeable. For example, a summary of statistics from one dataset should be easily mergeable with that of another dataset such that it is easy to summarize the statistics of the union of the two datasets. Similarly, it should be easy to transfer the learnings of one model to another model in general, and the learnings of a previous version of a model to the next version of the same model in particular.

There is however a catch, which relates to the previous discussion regarding the equivalents of test coverage for models. Merging new fragments into a model could necessitate creation of novel out-of-distribution and counterfactual evaluation data, contributing to the difficulty of (efficient) model evolution, thus rendering it a lot harder than pure code evolution.

https://blog.tensorflow.org/2020/09/brief-history-of-tensorflow-extended-tfx.html

References

[Callen] H.B. Callen "Thermodynamics and an introduction to thermostatistics: 2nd edition" John Wiley & Sons 1985.

[W] S. Watanabe "Algebraic geometry and statistical learning"

[MDLA1] MDLA "Deep learning is singular, and that's good" 2020.

[WAIC] S. Watanabe "A widely applicable Bayesian information criterion" JMLR 2013.

[H] J. Hestness et al "Deep learning scaling is predictable, empirically" arXiv: 1712.00409, 2017.

[KM] J. Kaplan, S. McCandlish et al "Scaling laws for neural language models" arXiv: 2001.08361, 2020.

[HKK] T. Henighan, J. Kaplan, M. Katz et al "Scaling laws for autoregressive generative modeling" arXiv: 2010.14701, 2020.

[A] D. Arovas "Lecture notes on thermodynamics and statistical mechanics"

[PLOW] M. Prokopenko, J.T. Lizier, O. Obst, X.R. Wang "Relating Fisher information to order parameters" Physical Review E 84 2011.