DLT Lecture 3 : Phase transitions

In a fint order phase transition the temperature or pressure are varied, theveby changing the Gibbs potential in such a way that two local minima (whose "identity" and separation are presewed by the variation) switch roles as the global minima. Each minima is a phase, and the different material propevties of the phase include diffewing values of molav enevgy, entwpy and other parametew [Callen, § 9]. In general we may use any thermodynamic potential in place of the Gibbs potential and a family of intensive parametew in place of temperature and pressure, but the essential point is that fint-order phase transitions are about the configuration of critical values in $\mathbb{R}$ [Callen, p. 220].

In a second order phase transition the "moduli" of the thermodynamic potential G (e.g. T, P) are varied so that as a function of the extensive coordinate (e.g. V) the potential has a degenerate critical point (typically the result of multiple minima "colliding" as $(T, P) \longrightarrow (T_{cr}, P_{cr})$). Thus a second-order phase transition is about the configuration of critical points in the space of moduli.

Following Watanabe we focused in (DLT2) on the free energy

$$F = F(n,\beta) = -\mathbb{E}\left[\log \int dw\, \varphi(w)\, e^{-n\beta L_n(w)}\right] = -\mathbb{E}\left[\log \int dw\, e^{-\beta H_n(w)}\right] \tag{1.1}$$

where $H_n(w) = nL_n(w) - 1/\beta \log \varphi(w)$, as our thermodynamic potential. Here $\mathbb{E}[-]$ denotes the expectation with respect to the dataset D_n. The proposed intensive coordinates are n, β thought of as inveve temperatures. But there is an obvious problem: what is the analogue of the extensive coordinate V?

Watanabe discusses phase transitions in [W, Remark 6.17] ($\beta \to \infty$), [W, Table 8.2] (which suggests the existence of phase transitions as a point of difference between regular and singular models) and in more detail in [W2, §9.4] and the lectures [W3]. The above issue is sidestepped (perhaps correctly) by defining a "phase transition" to be an infinitesimal curve in moduli space along which the posterior changes "drastically" [W2, Def^n 29], for sufficiently large n. This is a subtle and imprecise definition, so we adopt a more pedestrian approach.

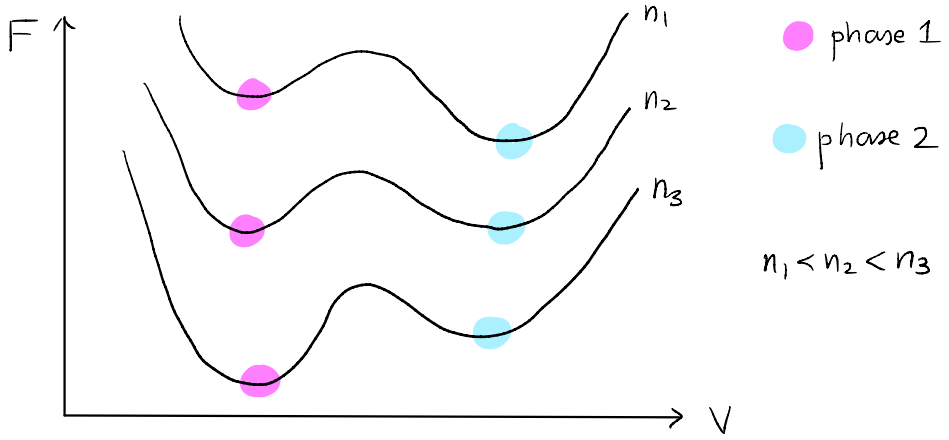
Suppose given an analytic function $V : W \to \mathbb{R}$. We define

$$F(n, \beta, V) = -\mathbb{E}\left[\log \int_{\{w \mid V(w) = V\}} dw\, \varphi(w)\, e^{-n\beta L_n(w)} \right] \qquad (2.1)$$

We can introduce a similar V-dependence to the average energy U and entropy S. We define a phase to be a critical point of $F(n, \beta, V)$ at fixed n, β as a function of V

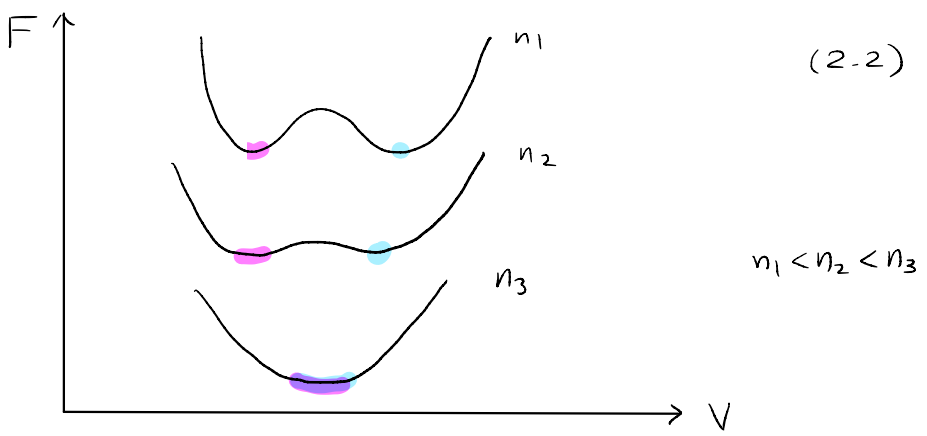
First-order phase transition in n at fixed β.

(cf [Callen, Fig. 9.4])



Second-order phase transition in n at fixed β

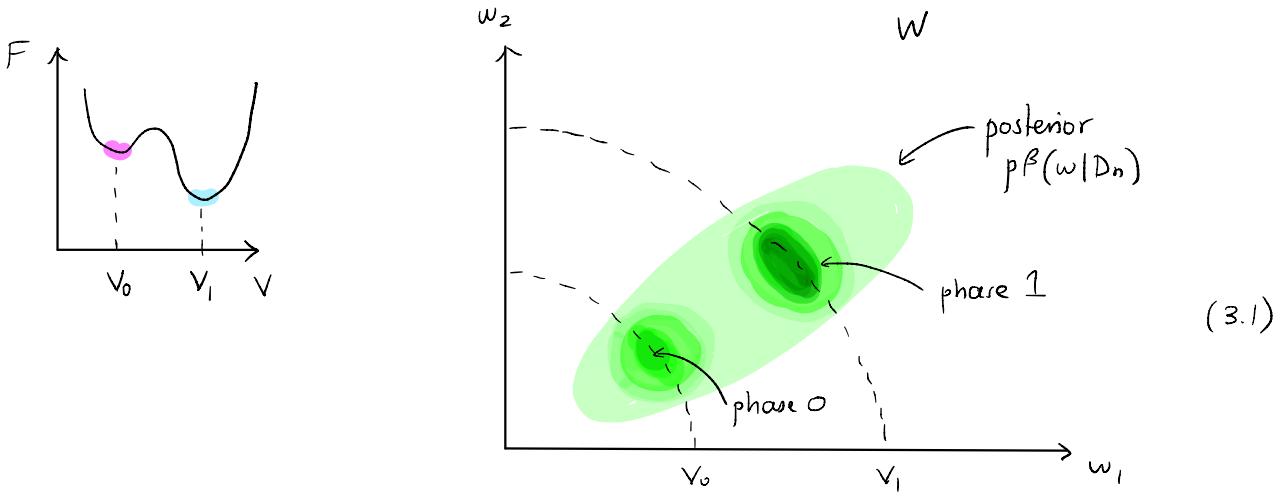
(cf [Callen, Fig. 9.6])



(2.2)

1. Studying the free energy

Suppose $V = \|-\|$ is the norm, and $W \subseteq \mathbb{R}^2$ with coordinates w_1, w_2.



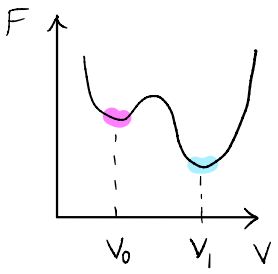
(3.1)

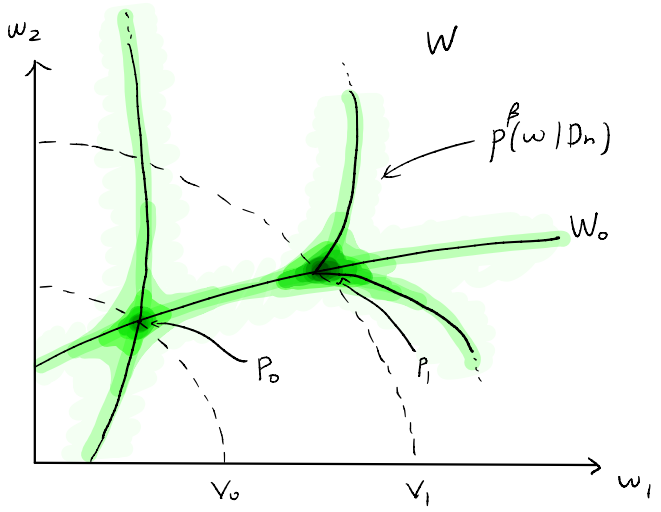
The minima of the free energy correspond to regions where the posterior is concentrated, as detected by the projection V. Varying the parameters n, β will vary the posterior distribution (e.g. the location of its maxima in W) and generically this will also cause variations in F reflecting these changes in the posterior. Hence we use the geometry of F as a proxy for the behaviour of the posterior.

Note that the average energy and entropy may differ between phases, in the sense that averaging over the posterior near the shown regions in W (or more precisely a range of level sets near V_0, V_1) may give different values. For instance the function K may have "more singular" critical points near $\{w \mid V(w) = V_1\}$ than near $\{w \mid V(w) = V_0\}$ (indeed we expect this). One also distinguishes phases by different values of higher derivatives of thermodynamic potentials.

A reasonable intuition is that each phase of the triple (P, Q, S) with respect to the chosen V is a <u>different kind of candidate solution</u> to the problem of modelling the true distribution.

The most natural examples of phases, or local minima of the free energy, are those associated to singularities of K (meaning a critical point $w \in W$ where $K(w) = 0$). An example of two phases associated to singularities P_0, P_1 is shown below:





(4.1)

However it is important to note that <u>not all phases are associated with points of W_0</u>. In general we expect a phase associated to any local minima of K in the following sense:

<u>Hypothesis</u> suppose that $w_0 \in W$ has the property that for ε sufficiently small, $(V_0 = V(w_0))$

$$K(w_0) = \inf\{ K(w) \mid V_0 - \varepsilon < V(w) < V_0 + \varepsilon \} \tag{4.2}$$

Let $\mathcal{E}(V) = \exp(-F(V))$ and note that (ignoring $\mathbb{E}[\cdot]$)

$$\begin{aligned} \mathcal{E}(V_0) &= \frac{d}{dU}\Big[\int_{V_0-\varepsilon}^{U} \mathcal{E}(V)\, dV \Big]\Big|_{U=V_0} \\ &= \frac{d}{dU}\Big[\int_{Q_U} dw\, \varphi(w)\, e^{-n\beta L_n(w)} \Big]\Big|_{U=V_0} \end{aligned} \tag{4.3}$$

where $Q_U = \{ w \in W \mid V_0 - \varepsilon < V(w) < U \}$. Since Q_U is semi-analytic the asymptotic methods of [W] apply (see e.g. [W, § 7.6] although this is not sufficiently careful, and we are also punting on some details).

Applying the resolution procedure to $K - K(w_0)$ on Q_U yields

(5.1)

$$\int_{Q_U} dw\, \varphi(w)\, e^{-n\beta L_n(w)} = \int_{Q_U} dw\, \varphi(w)\, e^{-n\beta(L_n(w)-L_n(w_0))} e^{-n\beta L_n(w_0)}$$

$$= e^{-n\beta L_n(w_0)} \int_{Q_U} dw\, \varphi(w)\, e^{-n\beta(K_n(w)-K_n(w_0))}$$

$$= e^{-n\beta L_n(w_0)} \frac{(\log n)^{m-1}}{n^\lambda} \int_{Q_U} du^* \int_0^\infty dt\, S(\beta, u)$$

where λ is the RLCT of $\{w \mid K(w) = K(w_0)\} \cap Q_U$ (this may be associated to a point on the level set $\{w \mid V(w) = V_0\}$ other than w_0, so be cautious!), and $S(\beta, u) = t^{\lambda-1} e^{-\beta t + \beta\sqrt{t}\,\xi(u)}$ where ξ describes fluctuations. Hence

$$\mathcal{E}(V_0) = e^{-n\beta L_n(w_0)} \frac{(\log n)^{m-1}}{n^\lambda} \frac{d}{dU}\left[\int_{Q_U} du^* \int_0^\infty dt\, S(\beta, u) \right]\Big|_{U=V_0} \quad (5.2)$$

and so

$$F(V_0) = n\beta L_n(w_0) + \lambda \log n - (m-1)\log\log n \quad (5.3)$$

$$- \log\left(\frac{d}{dU}\left[\int_{Q_U} du^* \int_0^\infty dt\, S(\beta, u) \right]\Big|_{U=V_0} \right)$$

Note that the final term depends on β but not n. This is perhaps an indication that phase transitions involving a variation in β are more subtle. For the moment we assume $\beta = \beta_0$ is fixed, and n large enough so that we may neglect the loglog term, so that as an approximation

$$F(V_0) \approx \underbrace{n\beta_0 L_n(w_0)}_{\text{energy}} + \underbrace{\lambda \log n}_{\text{entropy}} + \text{const.} \quad (5.4)$$

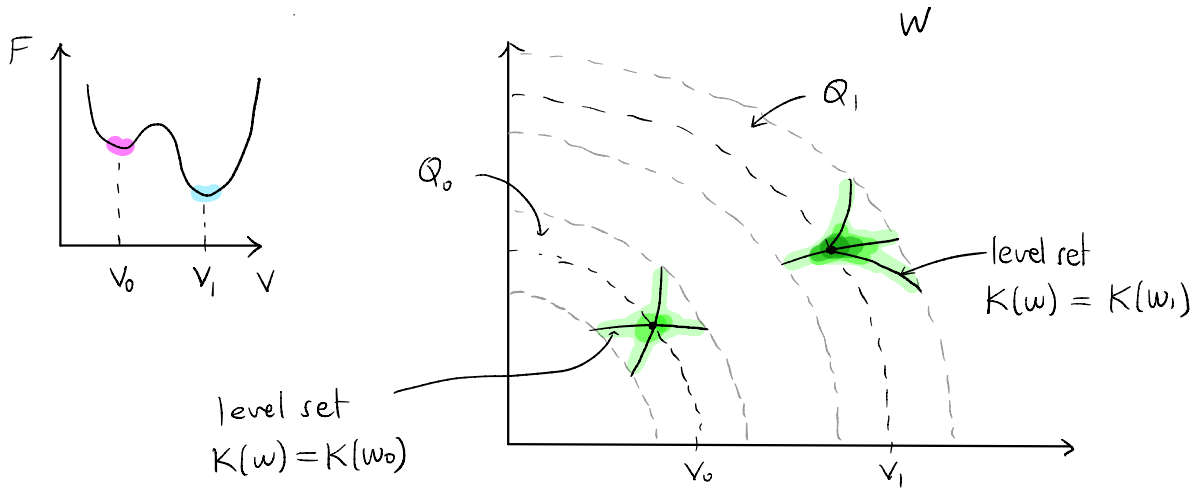
Since the RLCT λ of the level set $\{w \mid K(w) = K(w_0)\} \cap Q_U$ is rational, V_0 is a local minima of the free energy. Thus associated to any local minima of K in the above precise sense is a local minima of the free energy, that is, a phase.

Some of these phases will be associated, as in (4.1), with global minima of K but the learning process is governed by first-order phase transitions involving non-global local minima, as we describe in the next section.

In summary: classification of phases of the learning machine is closely related to the classification (in the usual sense of singularity theory) of critical points of K (or what is the same, singularities of level sets of K).

2. First-order phase transitions

There can be no first-order phase transition involving two phases associated to singularities of W_0 as in (4.1), because varying n cannot reverse the inequality between their free energies, which by (5.4) are determined solely by their local RLCT. So consider the following more general situation



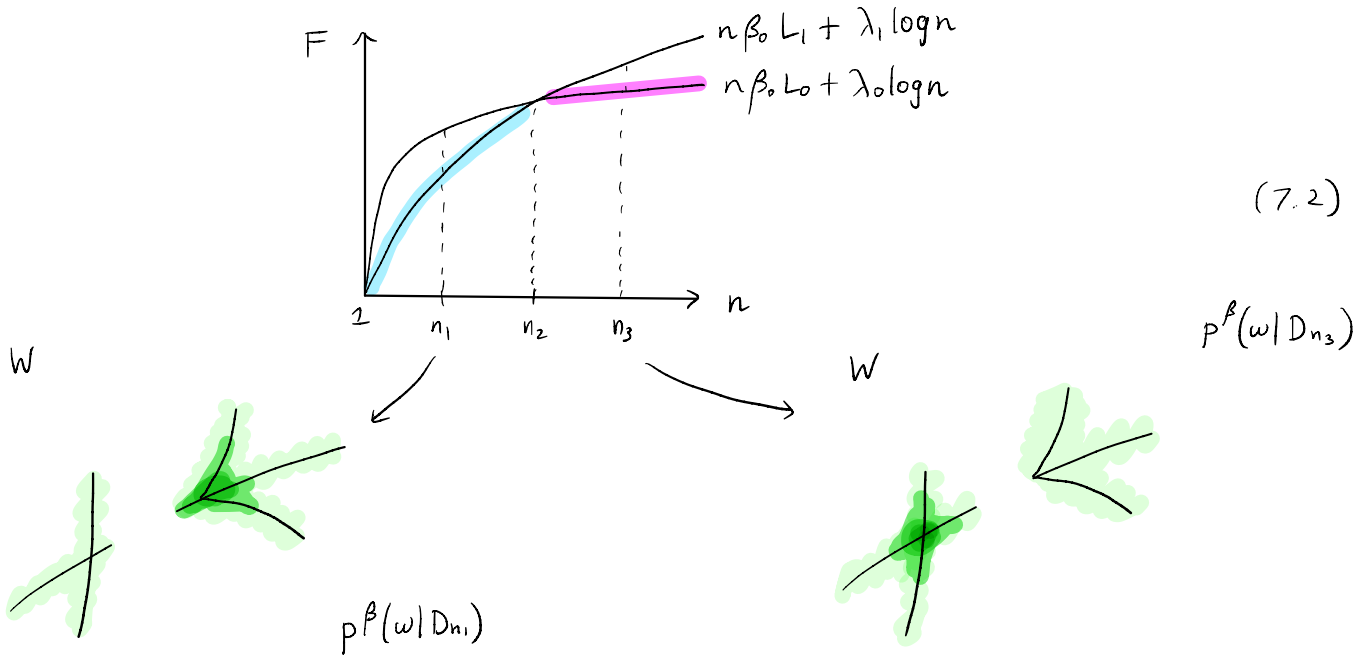
(6.1)

where $w_0, w_1 \in W$ are two local minima of K (w.r.t. V in the sense of (4.2)), $V_i = K(w_i)$ for $i \in \{0,1\}$ and $Q_i = \{w \in W \mid V_i - \varepsilon < V(w) < V_i + \varepsilon\}$ for some small ε.

We suppose $K(w_0) < K(w_1)$ but $\lambda_1 < \lambda_0$ where λ_i denotes the RLCT of the level set $\{w \mid K(w) = K(w_i)\}$. We claim that this implies a first-order phase transition, where phase 1 is preferred by the posterior for small n and phase 0 is preferred for large n. We assume n is large enough so $L_n(w_i) \approx L_i$ is approximately independent of n with $L_0 < L_1$ (this can be made precise using $K_n \to K$) so that (up to a constant)

$$F(V_i) \approx n\beta_0 L_i + \lambda_i \log n \tag{7.1}$$

Then by graphing the free energy as a function of n we can see the possibility of a first-order phase transition (at n_2 in the diagram):



(7.2)

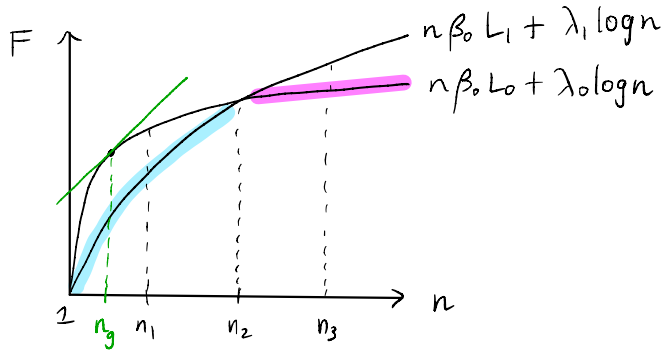
For small n, the phase ("candidate solution") which is lower entropy may be preferred by the posterior even if it has higher error, but as $n \to \infty$ the constraint on error comes to dominate ("at" $n = \infty$ the posterior concentrates at the point of lowest entropy on W_0).

Remark As far as I know the only discussion of first-order phase transitions in SLT is in [W3]. The above is largely based on that discussion.

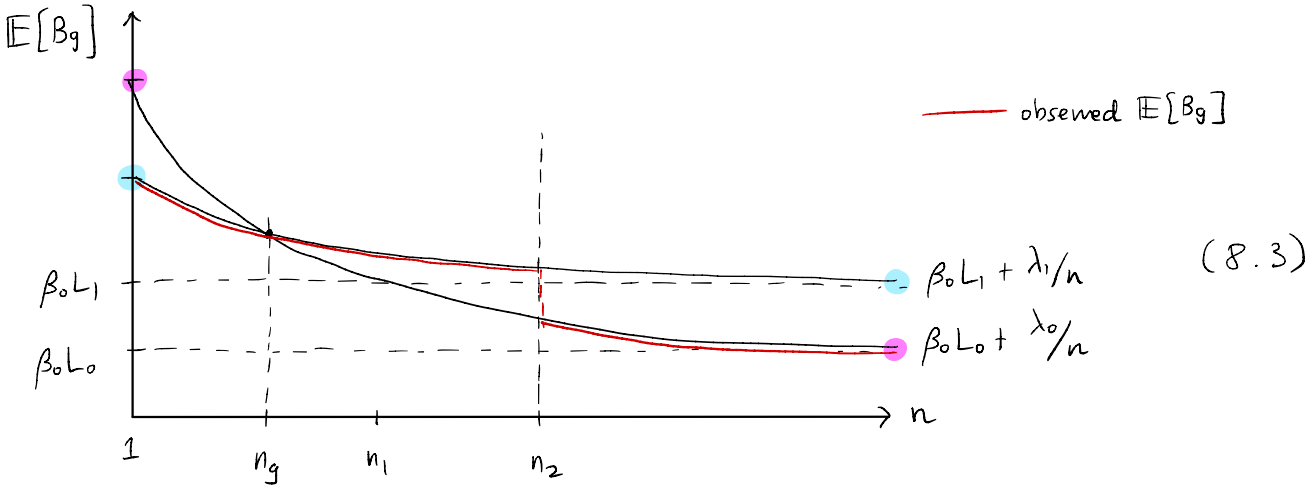
Remark Recall from DLT2 p.14 that we should think of the Bayes generalisation error as $\frac{\partial}{\partial n} F$, the extensive thermodynamic coordinate conjugate to the "inverse temperature" n. As in (2.1) we can add a V-dependence to the generalisation error, and from (7.1) we find that the generalisation error associated to a phase is (at least formally)

$$B_g(n, \beta, V) = \frac{\partial}{\partial n} F(V_i) \approx \beta_0 L_i + \frac{\lambda_i}{n} \qquad (8.1)$$

The case where $L_i = 0$ is covered by [W], we leave a rigorous discussion of the general case to elsewhere. Note that the preferred phase is selected to minimise (7.1) not (8.1), so the "full" $\mathbb{E}[B_g]$ will display a characteristic trace of the first-order phase transitions as shown below



(8.2)



(8.3)

Note the discontinuity in the observed generalisation error at the phase transition which is typical for extensive observables [Callen, p. 220] and corresponds to the discontinuity in the slope of F against n at the phase transition.

Problem 1 Demonstrate the existence of first-order phase transitions in small neural networks.

3. Second-order phase transitions

The learning machine of DLT2 §3 is determined by its Boltzmann distribution which varies with the parameter $\theta = (\theta^1, \theta^2) = (n, \beta)$. In the previous section we examined such variations via their effect on $F = F(V)$ as a proxy for a direct examination of the posterior. We now study the posterior directly. Observe that given an infinitesimal variation $\Delta\theta$, and writing $P(\theta) = p(w|D_n, \theta)$

$$P(\theta+\Delta\theta) \approx P(\theta) + \sum_i \Delta\theta^i \frac{\partial P}{\partial \theta^i}$$

$$\frac{P(\theta+\Delta\theta)}{P(\theta)} \approx 1 + \sum_i \frac{\Delta\theta^i}{P(\theta)} \frac{\partial P}{\partial \theta^i}$$

$$\ln\left(\frac{P(\theta+\Delta\theta)}{P(\theta)}\right) \approx \sum_i \frac{\Delta\theta^i}{P(\theta)} \frac{\partial P}{\partial \theta^i} - \frac{1}{2}\sum_{i,j} \frac{\Delta\theta^i \Delta\theta^j}{P(\theta)^2} \frac{\partial P}{\partial \theta^i} \frac{\partial P}{\partial \theta^j}.$$

Hence

$$KL(P(\theta+\Delta\theta) \,\|\, P(\theta)) \approx \int dw \left[P(\theta) + \sum_k \Delta\theta^k \frac{\partial P}{\partial \theta^k}\right] \cdot \left\{\sum_i \frac{\Delta\theta^i}{P(\theta)} \frac{\partial P}{\partial \theta^i} - \frac{1}{2}\sum_{i,j} \frac{\Delta\theta^i \Delta\theta^j}{P(\theta)^2} \frac{\partial P}{\partial \theta^i} \frac{\partial P}{\partial \theta^j}\right\}$$

$$= \sum_i \Delta\theta^i \int dw \frac{\partial P}{\partial \theta^i} - \frac{1}{2}\sum_{i,j} \Delta\theta^i \Delta\theta^j \int dw \frac{1}{P(\theta)} \frac{\partial P}{\partial \theta^i} \frac{\partial P}{\partial \theta^j} + \sum_{i,k} \Delta\theta^i \Delta\theta^k \int dw \frac{1}{P(\theta)} \frac{\partial P}{\partial \theta^i} \frac{\partial P}{\partial \theta^k} + O(\Delta\theta^3)$$

But $\int dw \frac{\partial P}{\partial \theta^i} = \frac{\partial}{\partial \theta^i} \int dw P = \frac{\partial}{\partial \theta^i}(1) = 0$ so

(10.1)

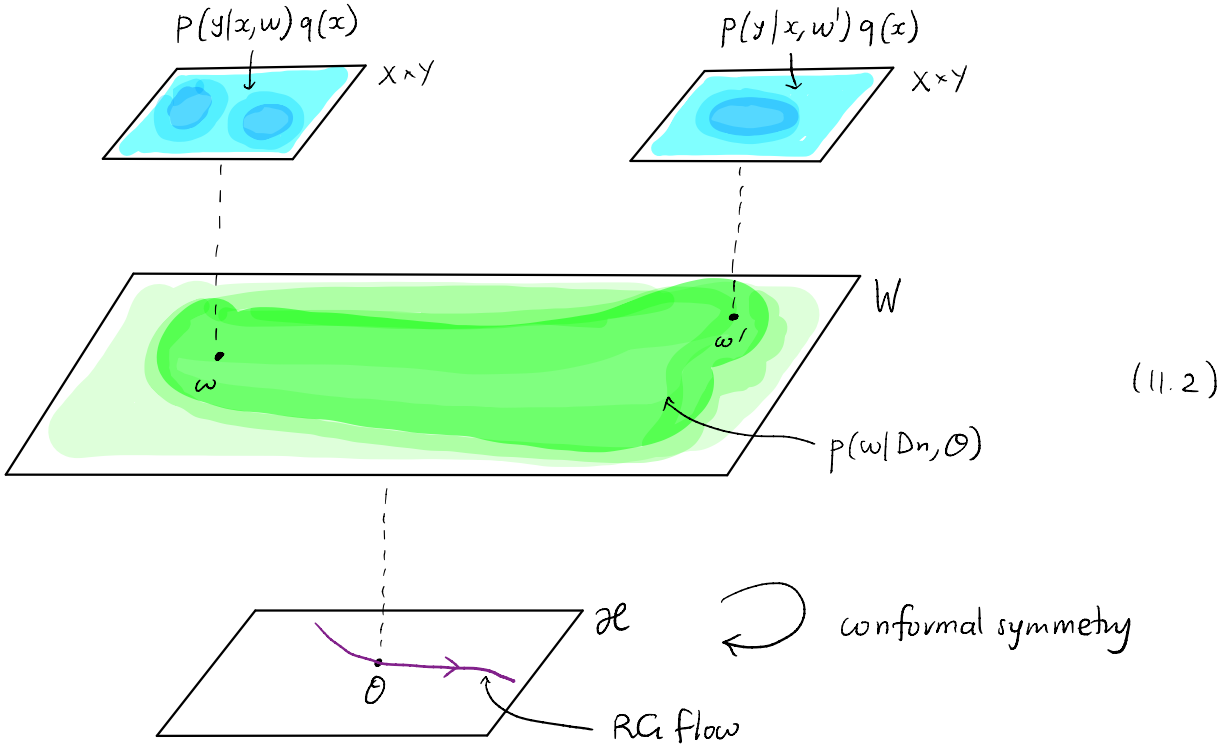
$$KL(P(\theta+\Delta\theta) \,\|\, P(\theta)) \approx \frac{1}{2}\sum_{i,j} \Delta\theta^i \Delta\theta^j \int dw \frac{\partial \log P(\theta)}{\partial \theta^i} \frac{\partial \log P(\theta)}{\partial \theta^j} P(\theta)$$

This means that if we want to understand how variations in θ affect the posterior, we must begin with the Fisher information matrix, which is a symmetric 2-form on the space $\mathcal{H}$ of parameters θ

$$F_{ij}(\theta) = \int dw\, P(\theta)\, \frac{\partial \log P(\theta)}{\partial \theta^i}\, \frac{\partial \log P(\theta)}{\partial \theta^j} \qquad (11.1)$$

We note that a priori this integral may not exist for all values of θ, and even if the matrix $F(\theta) = (F_{ij}(\theta))_{i,j}$ exists it may not be positive definite. It is the former phenomenon that is of interest in connection with second-order phase transitions, because if $F_{ij}(\theta) \longrightarrow \infty$ as $\theta \longrightarrow \theta_{cr}$ then small variations in n, β near n_{cr}, β_{cr} have divergent effects on the posterior (and thus on any quantity produced by integration over the posterior). I learned this information-theoretic point of view from [DFL], [PLOW].

Remark It is important to distinguish the Fisher matrix on $\mathcal{H}$ above from the Fisher matrix on W, considered for example in [W].



(11.2)

Remark $F_{ij}(\theta)$ is actually a random matrix due to the stochasticity in D_n.

Now recall that (of course this makes no sense currently for $\theta^i = n$)

$$\frac{\partial}{\partial\theta^i} F = -\frac{\partial}{\partial\theta^i}\log Z = \frac{-1}{Z}\frac{\partial}{\partial\theta^i}\int dw\, \exp(-n\beta L_n(w))\,\varphi(w)$$

$$= -\frac{1}{Z}\int dw\, \frac{\partial}{\partial\theta^i}\left[-n\beta L_n(w)\right] \exp(-n\beta L_n(w))\,\varphi(w) \qquad (12.1)$$

$$= \mathbb{E}_w^\beta\left[\frac{\partial}{\partial\theta^i}\left\{n\beta L_n(w)\right\}\right]$$

Hence, if we write $X^i = \frac{\partial}{\partial\theta^i}\left(n\beta L_n(w)\right)$,

$$\frac{\partial}{\partial\theta^i}\log P(\theta) = \frac{\partial}{\partial\theta^i}\log\left[\frac{1}{Z}\exp(-n\beta L_n(w))\,\varphi(w)\right]$$

$$= \frac{\partial}{\partial\theta^i}\left[-n\beta L_n(w)\right] - \frac{\partial}{\partial\theta^i}\log Z \qquad (12.2)$$

$$= -\left\{X^i - \mathbb{E}_w^\beta[X^i]\right\}$$

Hence

$$F_{ij}(\theta) = \int dw\left(X^i - \mathbb{E}_w^\beta[X^i]\right)\left(X^j - \mathbb{E}_w^\beta[X^j]\right)P(\theta) \qquad (12.3)$$

is the covariance matrix of the random variables X^1, X^2 (hence in particular positive semi-definite). The meaning of X^2 is straightforward [WAIC, (9)]

$$X^2 = nL_n(w) \qquad \text{"error"} \qquad (12.4)$$

$$\mathbb{E}_w^\beta[X^2] = nL_n(w_0) + nG_t \qquad \text{"Gibbs training error"}$$

The $\frac{\partial}{\partial n}$ derivative is more subtle, and we are tempted to set

$$X^1 = (n+1)\beta L_{n+1}(w) - n\beta L_n(w) \tag{13.1}$$

except that as it stands this does not make sense. However it seems we should associate $\mathbb{E}_w^\beta[X^1]$ to the Bayes training error, so provisionally we think of X^1 as a local density for the Bayes training error.

By Appendix A we have

$$F_{ij}(\theta) = \frac{\partial^2 F}{\partial\theta_i\partial\theta_j} + \mathbb{E}_w^\beta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\{n\beta L_n(w)\}\right] \tag{13.2}$$

and hence

$$F_{11}(\theta) = \frac{\partial^2 F}{\partial n^2} + \mathbb{E}_w^\beta\left[\frac{\partial}{\partial n}X^1\right]$$

$$F_{12}(\theta) = F_{21}(\theta) = \frac{\partial^2 F}{\partial n\partial\beta} + \mathbb{E}_w^\beta\left[\frac{\partial}{\partial n}(nL_n)\right] \tag{13.3}$$

$$F_{22}(\theta) = \frac{\partial^2 F}{\partial\beta^2}$$

Hence using (12.1), (12.4)

$$\boxed{F_{22}(\theta) = \frac{\partial}{\partial\beta}\mathbb{E}_w^\beta[X^2] = n\frac{\partial}{\partial\beta}G_t(n,\beta)} \tag{13.4}$$

Returning now to the topic of second-order phase transitions, if some component of F diverges as θ approaches some limiting value, then there are "effectively infinitely many" statistically distinct posterior distributions along a finite path in $\mathcal{H}$ (see [DFL, §III]). In the next section we sketch, using the renormalisation group, what the existence of such a phase transition suggests for the Gibbs training error.

4. Power laws and the renormalisation group

This section is speculative, and its aim is to present a possible path towards deriving power laws analogous to those in DLT1 within SLT. There is some overlap with [W4] but this remains to be understood. For an introduction to renormalisation see [F].

The metric tensor $\mathcal{F}$ on $\mathcal{H}$ is a highly nontrivial object, but it possesses some natural symmetries from which we can aim to extract scaling laws. The symmetries in question arise from the fact that <u>there is no canonical scale</u> for n, β and thus these parameters may be "rescaled" without "essentially changing" the statistics (up to a scale).

More precisely let $\mathcal{H}^*$ be the open subset of $\mathcal{H}$ where $\Theta \neq 0$ and $\mathcal{F}$ is positive-definite, so $\mathcal{H}^*$ is a Riemannian manifold, with coordinates

↑ associated to d below

$$J = 1/n, \quad T = 1/\beta, \quad \Theta = (J, T). \tag{14.1}$$

The Fisher information is the following 2-form on $\mathcal{H}^*$

$$\mathcal{F}(n,\beta)\, dn d\beta = \mathcal{F}(J,T) \frac{\partial n}{\partial T} dT \frac{\partial \beta}{\partial T} dT$$

$$= \frac{1}{J^2 T^2} \mathcal{F}(J,T)\, dJ dT$$

We write $\mathcal{F}^* = \frac{1}{J^2 T^2} \mathcal{F}$ so that in particular $\mathcal{F}^*_{22} = \frac{1}{J^3 T^2} \frac{\partial}{\partial \beta} G_t(J,T)$.

Suppose that $\mathcal{H}^*$ admits a conformal symmetry associated to the aforementioned rescaling, with infinitesimal generator

$$J \longrightarrow J + \epsilon K^1(\Theta) + O(\epsilon^2)$$
$$T \longrightarrow T + \epsilon K^2(\Theta) + O(\epsilon^2) \tag{14.2}$$

where $K^1 = AJ$, $K^2 = BT$, for constants A, B. Here we follow [DFL]. If we assume that there is an "intrinsic dimension" d as in [DFL, (5.2)] and that there is a fixed point of the RG flow at $(0,1)$ then this implies a system of differential equations for $F^*_{11}, F^*_{12}, F^*_{22}$ on $\mathcal{H}^*$. In particular

$$F^*_{22} K^2_{,2} + F^*_{22} K^2_{,2} + F^*_{22,r} K^r + d F^*_{22} = 0 \qquad (15.1)$$

where commas in subscripts denote derivatives as in [DFL, (5.2)]. This yields

$$2B F^*_{22} + AJ \frac{\partial}{\partial J}(F^*_{22}) + BT \frac{\partial}{\partial T}(F^*_{22}) = -d F^*_{22}$$

$$AJ \frac{\partial}{\partial J}(F^*_{22}) + BT \frac{\partial}{\partial T}(F^*_{22}) = (-d-2B) F^*_{22}$$

which suggests F^*_{22} is quasi-homogeneous: supposing $F^*_{22} = T^u T^v$ we obtain $uAJ^u T^v + vB J^u T^v = (-d-2B) J^u T^v$ hence

$$uA + vB = -d-2B \qquad (15.2)$$

From this we obtain for $(J,T) \approx (0,1)$

$$J^{-3} T^{-2} \frac{\partial}{\partial \beta} G_t \sim J^u T^v$$

$$\frac{\partial}{\partial \beta} G_t \sim J^{u+3} T^{v+2}$$

$$G_t \sim \rho J^{u+3} T^v + \text{const} \qquad (\rho \text{ constant})$$

and hence as $n \to \infty$

$$\boxed{G_t(n,1) \sim \frac{\rho}{n^{u+3}} + \text{const.}} \qquad (15.3)$$

Under many optimistic hypotheses and ignoring several mathematical issues this gives a "derivation" of power law behaviour for the Gibbs training error as a function of the dataset size n. The relation of the Gibbs generalisation error to deep learning practice (modulo the relation of SGD to the posterior) was pointed out in SLT6 p.⑦.

Turning this around, one could read the existence of power law behaviour in Transformer models as <u>evidence for second-order phase transitions</u> in the associated learning machines.

<u>Remarks</u> Here are some thoughts on turning this into real mathematics:

(i) One thing that is initially confusing is that in a second-order phase transition usually some generalised susceptibility <u>diverges</u> at the critical temperature. Since $J^{-3}T^{-2}\frac{\partial}{\partial\beta}G_t$ appear in $\mathcal{F}^*$ as long as this quantity diverges more slowly than J^{-3} (i.e. u, which is negative, satisfies $|u|<3$) as $J\to 0$ we have <u>convergence</u> of the Gibbs training error. It is of course not clear how to actually prove $|u|<3$.

(ii) The sketch of renormalisation group methods above assumes that the RG transformation is especially simple (i.e. that the flows "stay in $\mathcal{H}$"). This is unlikely. It seems that a proper treatment must reconcile RG flows in the information geometry setting with resolutions of singularities.

(iii) A fundamental approach to treating n as continuous must be found, as at the moment this prevents much of this lecture from being formalised.

(iv) In the realisable case by [W, Theorem 6.10] we have $n G_t \longrightarrow G_t^*$ and $\mathbb{E}[G_t^*] = \frac{\lambda}{\beta} - \nu(\beta)$ so we expect

$$G_t(n,1) \sim \frac{\lambda}{n} - \nu(1) \qquad (17.1)$$

which fits (15.3) with $u = -2$. In [W4] Watanabe shows that "conditions of learnability of index γ" imply that [W4, (19)]

$$G_t(n,1) \sim \frac{\rho}{n^\gamma} + \text{const.} \qquad (17.2)$$

for some constant ρ. The topic of renormalisation is discussed from a different point of view in [W4, §6.2].

(v) Note that we also expect power laws for the other components of $\mathcal{F}$, see (13.3).

(vi) It is probably unrealistic to expect to <u>prove</u> the existence of second-order phase transitions in a nontrivial deep neural network, and to rigorously derive the scaling exponents. But it does seem tractable to construct a mathematical framework in which the RG flow is well-defined, and prove that <u>if</u> a phase transition exists then certain relations hold among the exponents etc. The yoga of universality classes, relevant operators etc., would also be valuable in clarifying the overall structure of a mathematical theory of deep learning.

(vii) The treatment of the stochasticity in D_n above is inconsistent, and this is another area where some nontrivial effort seems to be necessary.

5. Conclusion

This brings us to the end of the invitation to the theory of deep learning, consisting of DLT1, DLT2 and this document. As outlined in DLT1 the discovery of power laws portends the emergence of deep learning as a general purpose technology. The mathematical theory behind this technology is both beautiful and likely to have profound impacts in the real world as the scale of deep learning systems increases (think factories or chemical plants).

Some parts of this theory are already clear, while others remain to be uncovered and are presently only visible in outline. In DLT2 we gave an introduction to Watanabe's singular learning theory as a thermodynamics of deep learning, and in this lecture we discussed first-order phase transitions and, more speculatively, second-order phase transitions which may (with work) give a theoretical basis to the aforementioned empirically observed power laws.

It is truly remarkable that resolution of singularities, one of the deepest results in algebraic geometry, together with the theory of critical phenomena and the renormalisation group, some of the deepest ideas in physics, are both implicated in the emerging mathematical theory of deep learning. This is perhaps a hint of the fundamental structure of intelligence, both artificial and natural. There is much to be done!

Appendix A

Let $f = f(w, \theta)$ and $Z = \int dw\, e^{-f}$. Then (writing $\mathbb{E}[h] = \frac{1}{Z}\int dw\, h\, e^{-f}$)

$$\frac{\partial}{\partial \theta_i} Z = -\int dw\, \frac{\partial f}{\partial \theta_i} e^{-f}$$

$$\frac{\partial^2}{\partial \theta_j \partial \theta_i} Z = -\int dw\, \frac{\partial^2 f}{\partial \theta_j \partial \theta_i} e^{-f} + \int dw\, \frac{\partial f}{\partial \theta_i} \frac{\partial f}{\partial \theta_j} e^{-f}$$

$$\frac{\partial^2}{\partial \theta_j \partial \theta_i} \log Z = \frac{\partial}{\partial \theta_j}\left(\frac{1}{Z}\frac{\partial Z}{\partial \theta_i}\right) = \frac{1}{Z^2}\left(\frac{\partial^2 Z}{\partial \theta_j \partial \theta_i} Z - \frac{\partial Z}{\partial \theta_i}\frac{\partial Z}{\partial \theta_j}\right)$$

$$= -\frac{1}{Z}\int dw\, \frac{\partial^2 f}{\partial \theta_j \partial \theta_i} e^{-f} + \frac{1}{Z}\int dw\, \frac{\partial f}{\partial \theta_i}\frac{\partial f}{\partial \theta_j} e^{-f} - \frac{1}{Z^2}\frac{\partial Z}{\partial \theta_i}\frac{\partial Z}{\partial \theta_j}$$

$$= -\mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_j \partial \theta_i}\right] + \mathbb{E}\left[\frac{\partial f}{\partial \theta_i}\frac{\partial f}{\partial \theta_j}\right] - \mathbb{E}\left[\frac{\partial f}{\partial \theta_i}\right]\mathbb{E}\left[\frac{\partial f}{\partial \theta_j}\right]$$

$$\mathbb{E}\left(\left(\frac{\partial f}{\partial \theta_i} - \mathbb{E}\left[\frac{\partial f}{\partial \theta_i}\right]\right)\left(\frac{\partial f}{\partial \theta_j} - \mathbb{E}\left[\frac{\partial f}{\partial \theta_j}\right]\right)\right)$$

$$= \mathbb{E}\left[\frac{\partial f}{\partial \theta_i}\frac{\partial f}{\partial \theta_j}\right] - \mathbb{E}\left[\frac{\partial f}{\partial \theta_i}\right]\mathbb{E}\left[\frac{\partial f}{\partial \theta_j}\right]$$

$$= \frac{\partial^2}{\partial \theta_j \partial \theta_i} \log Z + \mathbb{E}\left[\frac{\partial^2 f}{\partial \theta_j \partial \theta_i}\right]$$

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.

[W2] S. Watanabe "Mathematical Theory of Bayesian Statistics" CRC Press 2018.

[W3] S. Watanabe "Statistical Learning Theory 13 : Phase transition and Prior effect" 2020
http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/slt202013.pdf

[G] R. Gilmore "Catastrophe theory for scientists and engineers".

[DFL] L. Diósi, G. Forgács, B. Lukács "Metricization of thermodynamic state-space and the renormalization group" Phys. Review A 1984.

[W4] S. Watanabe "Asymptotic learning curve and renormalizable condition in statistical learning theory" 2010.

[F] M. E. Fisher "Renormalization group theory: its basis and formulation in statistical physics" Reviews of Modern Physics 1998.