Singular Learning Theory 13 : Asymptotics of the free energy

19/4/21

In this note we examine the derivation [W, p.31-33] of the asymptotic behaviour of the free energy of a singular model. The result is proven fully in [W, Main Theorem 6.2, p.174] and requires some substantial groundwork to establish rigorously. Our aim here is to give a modest elaboration on [W, p.31-33] emphasising some conceptual points.

We assume familiarity with the thermodynamic point of view on singular learning theory elaborated in DLT2. From this point of view a thermodynamic system is completely understood if we know its <u>fundamental relation</u> giving a thermodynamic potential (energy, entropy or one of the Legendre transforms such as the Helmholtz or Gibbs free energies) as a function of extensive or intensive parameters. This determines the behaviour of the system when an internal constraint is removed by the postulates [C, Postulate 2] of thermodynamics and the equivalent form for potentials other than the entropy. In DLT2 the analogy is made between the <u>free energy</u> F_n° of [W] and the Gibbs potential [DLT2, p.14], [DLT3] so that the appropriate principle is:

<u>Gibbs Potential Minimum Principle</u> The equilibrium value of any unconstrained internal parameter in a system in contact with a thermal and a pressure reservoir minimises the Gibbs potential at constant temperature and pressure (those of the reservoir).

It is important to keep in mind that the free energy is the logarithm of an integral over some set of states <u>consistent with a constraint</u> (a typical physical example is volume V) [C, §16.1] and thus the free energy becomes a function of the value of that constraint. We view $F = F_n^\circ$ as not only depending on n, β but also on one (or more) auxiliary analytic functions $V : W \to \mathbb{R}$ so that $F = F(n, \beta, V)$. Taking this V into account allows us to study <u>phases</u> and <u>phase transitions</u> [DLT3]. Hence our aim in this note is to justify [DLT3, p.5], i.e.

<u>Aim</u> : derive the asymptotic behaviour of $F(n, \beta, V)$ as $n \to \infty$.

1. Coarse-graining

From an information theoretic point of view the idea of a continuous parameter space W is a useful mathematical fiction. Since we can only ever gather finitely many bits of information about the generating process out in the world (say finitely many numbers to finite precision) we can only ever make finitely many distinctions between possible models. So in fact statistical learning is in practice not a matter of comparing points $w_1, w_2 \in W$ but in fact is about comparing compact subsets $W_1, W_2 \subseteq W$ (for example $W_i = \{w \in W \mid V(w) \in [i, i+1]\}$ a partition of W by the values of some "observable" V).

For models with a nondegenerate Fisher information metric the difference between comparing points and comparing compact sets is only superficial: any comparison between sets will reduce to a comparison between the local minima of the KL divergence they contain. However the difference is profound in the singular case, because any integral over W near a local minima of K will be strongly affected by singularities. Since these integrals are the only "real" quantities, this means that at the coarse-grained level at which we actually "perceive" W, the behaviour is dominated by singularities (which are themselves in some sense "imperceptible" since they live at the fine-grained level of points $w \in W$ which we cannot have direct knowledge of).

Assume given a triple $(p(x|w), q(x), \varphi(w))$ as usual, satisfying fundamental conditions (I), (II) with $s = 2$. Let W denote the space of parameters, and $K : W \longrightarrow \mathbb{R}$ the KL divergence, K_n the empirical estimate according to some sample D_n. The posterior probability of $w \in W$ is [DLT2, p. 4] $p(w|D_n)dw = \frac{1}{Z_n^\circ}\varphi(w)e^{-nK_n(w)}dw$ and hence for any real analytic $V : W \longrightarrow \mathbb{R}$, writing $\int_{a<V<b}$ for $\int_{V^{-1}([a,b])}$

$$\begin{aligned} \mathbb{P}(a<V<b) &= \int_{a<V<b} p(w|D_n)dw \\ &= \frac{1}{Z_n^\circ}\int_{a<V<b} \varphi(w)e^{-nK_n(w)}dw \end{aligned} \tag{2.1}$$

As explained in [DLT2, p.⑨] we view the tempered posterior $p^\beta(w|D_n)$ as the Boltzmann distribution for a thermostatistical system with random Hamiltonian

$$H_n(w) = nK_n(w) - \frac{1}{\beta}\log \varphi(w) \qquad (3.1)$$

so that $e^{-\beta H_n(w)} = \varphi(w)e^{-n\beta K_n(w)}$. Since W is compact we may set

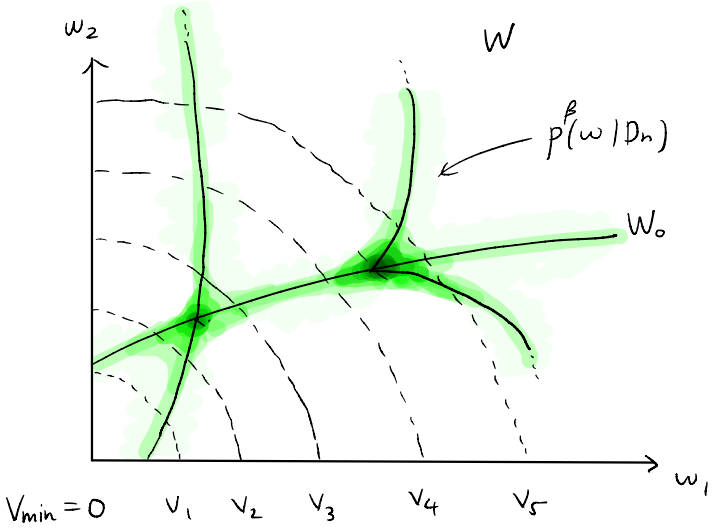
$$V_{min} = \inf\{V(w) \mid w \in W\}, \quad V_{max} = \{V(w) \mid w \in W\} \qquad (3.2)$$

and adopt for our fine-graining of W the "partition" of W into <u>cells</u> W_j determined by a partition of the interval $[V_{min}, V_{max}]$ as follows

$$V_{min} = a_0 < a_1 < \cdots < a_j < a_{j+1} < \cdots < a_{N+1} = V_{max} \qquad (3.3)$$

$$W_j = \{w \in W \mid a_j \leq V(w) \leq a_{j+1}\}$$

The following diagram from [DLT3, p.④] depicts an example where $W \subseteq \mathbb{R}^2$ and $V = \|-\|$



(3.4)

If we average the posterior $p^\beta(w|D_n)$ over each cell then we obtain a <u>coarse-graing</u> of the posterior associated to (3.3), which is a distribution over indices $j \in \{0, \ldots, N\}$.

$$p^\beta(j|D_n) = \int_{W_j} p^\beta(w|D_n)\,dw$$
$$= \frac{1}{Z_n^\circ}\int_{W_j} \varphi(w)\, e^{-n\beta K_n(w)}\,dw \qquad (4.1)$$

If we formulate this again as a Boltzmann distribution at inverse temperature β, with Hamiltonian $\mathcal{H}(j)$, then $p^\beta(j|D_n) = \frac{1}{Z} e^{-\beta \mathcal{H}(j)}$ and $Z = \sum_j e^{-\beta \mathcal{H}(j)}$ so

$$\frac{1}{Z} e^{-\beta \mathcal{H}(j)} = \frac{1}{Z_n^\circ}\int_{W_j} \varphi(w)\, e^{-n\beta K_n(w)}\,dw \qquad (4.2)$$

$$\mathcal{H}(j) = -\frac{1}{\beta}\log\left(\frac{Z}{Z_n^\circ}\int_{W_j} \varphi(w)\, e^{-n\beta K_n(w)}\,dw\right)$$

$$= -\frac{1}{\beta}\log\int_{W_j} \varphi(w)\, e^{-n\beta K_n(w)}\,dw + C$$

We refer to $\mathcal{H}(j)$ as the <u>coarse-grained Hamiltonian</u> and it governs (from the thermostatistical perspective) the probabilities of the "state" transitioning between different values of j. Of course up to the factor of $\frac{1}{\beta}$ this is precisely the <u>free energy of the compact set W_j</u>. Thus comparing the free energies of W_j and W_k (the standard method of model selection [WAIC]) is equivalent to <u>comparing the coarse-grained Hamiltonian</u> (and thus in a sense the statistician choosing a parameter in a sense <u>is</u> the dynamical system whose transitions are governed by $\mathcal{H}$).

Our goal is therefore to compute the asymptotics of integrals of the form

$$\int_{a<V<b} \varphi(w)\, e^{-n\beta K_n(w)}\,dw \qquad (4.3)$$

<u>Remark</u> For more on coarse-graining see [EL, p.483], [Ca, p.29], [F].

2. Asymptotics

Assume given a triple $(p(x|w), q(x), \varphi(w))$ as usual, satisfying fundamental conditions (I), (II) with $s = 2$. Recall that part of fundamental condition (I) is realisability. Given $V : W \longrightarrow \mathbb{R}$ analytic the set $\{w \in W \mid a \leq V \leq b\}$ can be cut out of W by two additional inequalities, and so [W, Main Theorem 6.2] applies just as well to the parameter space $V^{-1}([a,b])$ as it does to W provided the realisability condition still holds (i.e. $W_0 \cap V^{-1}([a,b]) \neq \phi$, and we may want to choose a, b to be regular values of V to be safe).

However, from the point of view of phase transitions we do not wish to assume this. For this reason we follow the derivation on [W, p. 31–33] as closely as we are able, up to the point where realisability is used.

References

[W] S. Watanabe "Algebraic geometry and statistical learning theory" 2009.

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

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

[Ca] J. Cardy "Scaling and Renormalization in Statistical Physics" Cambridge University Press 1996.

[EL] G. G. Emch, C. Liu "The logic of Thermostatistical Physics" Springer-Verlag 2002.

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