Notes on ‘Statistical Mechanics of Learning from Examples’

Rohan Hitchcock

28 September 2023

In this note we review the paper ‘Statistical Mechanics of Learning from Examples’ [SST92].

1 Set-up

Let $X \in \mathbb{R}^m$ be a random variable with unknown distribution $X \sim q(x)$, and consider a function $g : \mathbb{R}^m \to \mathbb{R}$ (also not known). Given examples $D_n = \{(X_1, g(X_1)), \ldots, (X_n, g(X_n))\}$, where the $X_i \sim q(x)$ are iid, our goal is to train a neural network to learn the function g.

Consider a neural network f with weights $W \in \mathbb{R}^d$. Let $\epsilon(X|W) \geq 0$ be a loss function for f measuring the deviation of $f(X|W)$ from $g(X)$. For example we could take quadratic loss:

$$\epsilon(X|W) = \tfrac{1}{2}(f(X|W) - g(X))^2 \ .$$

Given a loss function, we define the *training energy*

$$E(W) = \sum_{i=1}^{n} \epsilon(X_i|W)$$

and the *generalisation function*

$$\epsilon(W) \coloneqq \mathbf{E}_X[\epsilon(X|W)] = \int_{\mathbb{R}^m} \epsilon(x|W) q(x) dx \ .$$

Training the neural network using (full batch) gradient descent amounts to allowing the weights W to evolve according to (a discretisation of) the gradient flow of $E(W)$:

$$\frac{\partial W}{\partial t} = -\nabla_W E(W)$$

The paper considers a generalisation of this. We consider the weights W evolving according to the stochastic differential equation, called the *Langevin equation*, which is

$$\frac{\partial W}{\partial t} = -\nabla_W E(W) - \nabla_W V(W) + \eta(t) \tag{1}$$

where:

- $\eta(t)$ is “white noise”, where covariance of $\eta(t)$ and $\eta(t')$ is $2T\delta_{ij}\delta(t - t')$. Here $T \geq 0$ is the *temperature.*
- $V(W)$ represents “possible constraints on the range of weights”. In contrast to $E(W)$, $V(W)$ does *not* depend on the examples D_n.

Setting $V(W) = 0$ and $T = 0$ recovers gradient flow dynamics.

We consider the weights W evolving according to (1) in the limit as $t \to \infty$. To be more precise, we consider a solution W_t to the stochastic differential equation (1) — a stochastic process — and consider the random variable W obtained by taking the limit of W_t as $t \to \infty$. This requires some thought: firstly about the existence and uniqueness of solutions to (1), and secondly about the existence and sense in which the limit of W_t should be taken. The paper claims that the distribution of such a W is given by the density function on $\mathbb{R}^n$

$$p_\beta(w|D_n) = \frac{1}{Z_n^\beta} e^{-\beta E(w)} \qquad \text{where } Z_n^\beta = \int_{\mathbb{R}^d} e^{-\beta E(w)} \varphi(w) dw$$

where $\beta = 1/T$ and the effect of $V(W)$ is incorporated into a prior distribution $\varphi(w)$ on $\mathbb{R}^d$.

Definition 1. We define the following quantities as functions of the inverse temperature β and number of training examples n

- The *average training error*

$$\epsilon_t(\beta, n) = \frac{1}{n} \mathbf{E}_{D_n} \mathbf{E}_W [E(W)]$$

- The *generalisation error*

$$\epsilon_g(\beta, n) = \mathbf{E}_{D_n} \mathbf{E}_W [\epsilon(W)]$$

- The *free energy*

$$F(\beta, n) = \frac{-1}{\beta} \mathbf{E}_{D_n} [\log(Z_n)]$$

- The *entropy*

$$S(\beta, n) = -\mathbf{E}_{D_n} \int_{\mathbb{R}^d} \log(p_\beta(w)) p_\beta(w) \varphi(w) dw$$

Connection to singular learning theory

Singular learning theory [Wat09] considers learning in neural networks by considering the conditional distribution of an 'output' random variable $Y \in \mathcal{Y}$ given an 'input' random variable $X \in \mathcal{X}$. One studies a 'truth-model-prior' triplet $(q(y|x), p(y|x, w), \varphi(w))$, where:

- $q(y|x)$ is the (unknown) true conditional distribution of Y given X.
- $p(y|x, w)$ is the conditional distribution of Y given X modelled by the neural network with weights $w \in \mathcal{W}$.
- $\varphi(w)$ is a prior distribution on the weight space $\mathcal{W}$.

The central object of singular learning theory is the *Kullback-Leibler divergence* of $p(y|x, w)$ with respect to $q(y|x)$ considered as a function of w:

$$K(w) = \int_{\mathcal{X} \times \mathcal{Y}} \log \left(\frac{q(y|x)}{p(y|x, w)} \right) q(y|x) q(x) dx dy$$

Given a dataset $D_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$, one can also consider the so-called *empirical Kullback-Leibler divergence*

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} \log \left(\frac{q(Y_i|X_i)}{p(Y_i|X_i, w)} \right) .$$

One can write the *Bayes posterior distribution* at inverse temperature β in terms of $K_n(w)$ as

$$p_\beta(w|D_n) = \frac{1}{Z_n} e^{-\beta n K_n(w)} \varphi(w) dw \qquad \text{where } Z_n^\beta = \int_{\mathcal{W}} e^{-\beta n K_n(w)} \varphi(w) dw \ .$$

We define the *free energy* as $\widetilde{F}(\beta, n) = -\log(Z_n)$.

Theorem 2 ([Wat09, Main Theorem 6.2])**.** *Given some fairly general conditions, we have the asymptotic expansion (in probability) of the free energy*

$$\widetilde{F}(\beta, n) = \lambda \log n + (\mu - 1) \log \log n + O_p(1)$$

where λ and μ are geometric invariants of $K(w)$. Likewise the expected free energy has the asymptotic expansion

$$\mathbf{E}_{D_n}\left[\widetilde{F}(\beta, n)\right] = \lambda \log n + (\mu - 1) \log \log n + O(1) \ .$$

Returning to the setting of [SST92], suppose we take the loss function

$$\epsilon(X|W) = \tfrac{1}{2}(f(X|W) - g(X))^2 \ .$$

Given our network f, we define a statistical model

$$p(y|x, w) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(f(x|w)-y)^2}$$

and true distribution

$$q(y|x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-1}{2}(g(x)-y)^2} \ .$$

Lemma 3. *The Kullback-Leibler divergence of the model with respect to the truth is*

$$K(w) = \int_{\mathbb{R}^n} \tfrac{1}{2}(f(x|w) - g(x))^2 q(x) dx$$

and the empirical Kullback-Leibler divergence is

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} \tfrac{1}{2}(f(X_i|w) - g(X_i))^2 \ .$$

Proof. See [Car21, Lemma A.2]. □

It follows that:

- The training energy is $E(w) = nK_n(w)$.
- The limiting distribution of (1) $p_\beta(w|D_n)$ coincides with the Bayes posterior distribution, and likewise for Z_n^β.
- The free energy from [SST92] and [Wat09] are related by $F(\beta, n) = \frac{1}{\beta}\mathbf{E}_{D_n}\left[\widetilde{F}(\beta, n)\right]$. In particular we have the asymptotic expansion

Hence [Wat09, Main Theorem 6.2] gives us (under reasonable assumptions) the asymptotic expansion

$$F(\beta, n) = \frac{\lambda}{\beta} \log(n) + \frac{\mu - 1}{\beta} \log \log n + O(\tfrac{1}{\beta}) \ . \tag{2}$$

2 Statistical physics of learning

Recall that n is the number of training examples and d is the number of weights in the network. We define the ratio

$$\alpha = \frac{n}{d} \ .$$

The idea behind considering this ratio is that, according to principles from statistical mechanics, the training energy $E(W)$ should be *extensive*. This means that it should be proportional to the total degrees of freedom d. Since $\mathbf{E}_{D_n} E(W) = n\epsilon(W)$ this implies that[1] n should be proportional to d. This proportionality constant is α.

2.1 Techniques

This paper considers three methods of approximating the behaviour of the learning system: the high temperature limit (or more generally the high temperature expansion), the annealed approximation, and the replica method. Each of these methods essentially amounts to using different methods of approximating the free energy.

The high temperature limit The high temperature limit investigates the limiting behaviour of the system as

$$\alpha \to \infty \qquad \beta \to 0 \qquad \text{such that} \quad \alpha\beta < \infty$$

Under the high temperature limit it is shown that the distribution $p_\beta(w|D_n)$ approaches

$$p_0(w|D_n) = \frac{1}{Z^0} e^{-d\beta\alpha\cdot\epsilon(W)} \qquad \text{where } Z^0 = \int e^{-d\beta\alpha\epsilon(W)}\varphi(w)dw$$

That is, the training energy $E(W)$ is replaced by its average $E_0 := \mathbf{E}_{D_n}[E(W)] = n \cdot \epsilon(W) = \alpha d\epsilon(W)$. Notice that, if we are using the singular learning theory set up, $E_0 = nK(w)$ and the high temperature limiting distribution given in terms of the KL divergence

$$p_0(w|D_n) = \frac{1}{Z^0} e^{-n\beta K(w)} \qquad \text{where } Z^0 = \int e^{-n\beta K(w)}\varphi(w)dw$$

At times this has been referred to as the ‘annealed’ distribution.

More generally one can consider the *high temperature expansion*, which expands the free energy F as a power series in β of the form

$$-\beta F = \log Z^0 + \sum_{i=1}^{\infty} \beta^j F_j(\alpha\beta) \ .$$

The high temperature limit is recovered by approximating the free energy by the first term in the high temperature expansion.

The annealed approximation In the annealed approximation we define the *annealed free energy* as

$$F_{\text{an}} = \frac{-1}{\beta} \log\left(\mathbf{E}_{D_n}\left[Z_n^\beta\right]\right)$$

Since the logarithm is a convex function we have, by Jensen's inequality, that $F_{\text{an}} \leq F$. In the annealed approximation we replace the free energy F by the annealed free energy F_{an}. In [SST92] it is claimed that the annealed approximation works well in the following circumstances:

[1] Assuming $\epsilon(W)$ does not depend on d, which is frequently true in practice.

- In the limit as $\beta \to 0$, recovering the high temperature limit discussed above.
- In the limit as $\alpha \to \infty$ and finite β when the learning problem is realisable.

The annealed approximation can be understood as the exact theory for a certain system where both the weights and examples are both subject to dynamics:

$$\frac{\partial W}{\partial t} = -\nabla_W E(W) + \eta(t) \ , \qquad \frac{\partial X_i}{\partial t} = -\nabla_{X_i} E(W) + \eta_i(t) \ .$$

The replica method In this method the free energy is approximated by

$$-\beta F = \lim_{r \to 0} \frac{1}{r} \log \left(\mathbf{E}_{D_n} \left[Z^r \right] \right) \ .$$

Notice that "approximating" the above limit by taking $r = 1$ recovers the annealed free energy.

2.2 Results

Using these approximations, [SST92] derives *phase transitions* in the quantity α in various models. The most comprehensive and impressive results are for a network with discrete weights: $W_i \in \{-1, 1\}$ for all $1 \leq i \leq d$. Under the realisability assumption two critical values $\alpha_c(\beta)$ and $\alpha_o(\beta)$ are derived:

- For $\alpha > \alpha_c(\beta)$ perfect learning (learning with zero generalisation error) becomes possible.
- For $\alpha > \alpha_o(\beta)$ 'metastable' states vanish. Above this threshold perfect learning happens quickly.

These results remain valid even at low or zero temperature, and are verified with experiments. Phase diagrams in α and β are given. Similar results are derived for other networks, and without the realisability assumption.

Another result derived is that, for a "smooth network" (a network with real-valued weights and error function which is twice differentiable), the average training error and generalisation error are both $O(1/\alpha)$. This holds for both realisable and unrealisable learning problems.

Caveat Buried [SST92, Section VII.G] the assumption that the optimal network parameters are a discrete set in parameter space. We know that for networks with real-valued weights this is not the case [Wat09]. It is not immediately clear to me where these assumptions come into play. It is mentioned the authors plan to address relaxing this assumption in a future work.

References

[Car21] Liam Carroll. ‘Phase Transitions in Neural Networks’. MSc Thesis. The University of Melbourne, Oct. 2021.

[SST92] H. S. Seung, H. Sompolinsky and N. Tishby. ‘Statistical Mechanics of Learning from Examples’. In: *Physical review A* 45.8 (1992), p. 6056.

[Wat09] Sumio Watanabe. *Algebraic Geometry and Statistical Learning Theory*. Cambridge University Press, 2009.

Notation dictionary

Description	**Notation in this note**	**Notation in** [SST92]
Input random variable	$X \in \mathbb{R}^m$	$\mathbf{S} \in \mathbb{R}^M$
Input distribution	$q(x)$	$d\mu(\mathbf{S})$
Input dimension	m	M
Rule being learned	$g : \mathbb{R}^m \to \mathbb{R}$	$\sigma_0 : \mathbb{R}^M \to \mathbb{R}$
Number of examples	n	P
Neural network	$f(X\|W)$	$\sigma(\mathbf{W}; \mathbf{S})$
Weights of the neural network	$W \in \mathbb{R}^d$	$\mathbf{W} \in \mathbb{R}^N$
Number of weights	d	N
Loss function / error function	$\epsilon(X\|W)$	$\epsilon(\mathbf{W}; \mathbf{S})$
Expectation with respect to the training set	$\mathbf{E}_{D_n}$	$\langle\langle\ \rangle\rangle$
Expectation with respect to the weights	$\mathbf{E}_W$	$\langle\ \rangle_T$