Mean-Field Langevin Dynamics and Energy Landscape of Neural Networks

Pierre-Henri Guittard[1], Nicolas Richard[1], Fouad Doukkali[1], and Anja Miaro Raohilison[2]

[1]Master 2 IRFA
[2]Master 2 MMMEF

March 2020

1 Introduction

We try to review and explain the aforementioned article. We proceed by a thematic approach. Mean field theory focuses on the behavior of high-dimension stochastic models. The high-dimension term in our course relates to neural networks. The link between mean-field theory and neural networks is formed thanks to Langevin Dynamics. More precisely thanks to Langevin's equation role in non-convex optimisation.
This article covers a wide-range of subjects from statistical physics, analysis, optimization and variation calculus. Analyzing it gave us a hard time still we are satisfied with the result.

2 Well-posedness of the problem

2.1 Convex mixture of probabilities

We first prove that the potential function proposed in (1.3) is convex. In order to do this let us recall the following result:

Proposition 2.1. *Let m and m' be two probability measures on $(\Omega, \mathscr{E})$ a sigma-algebra endowed set. Let α in [0;1] then $\alpha m + (1-\alpha)m'$ is a probability on $(\Omega, \mathscr{E})$*

Proof. $(\alpha m + (1-\alpha)m')(\Omega) = \alpha m(\Omega) + (1-\alpha)m'(\Omega) = \alpha + (1-\alpha)$
Let A ⊂ B ⊂ E. It holds that:

$$\begin{aligned}(\alpha m + (1-\alpha)m')(B/A) &= \alpha m(B/A) + (1-\alpha)m'(B/A)\\ &= \alpha[m(B) - m(A)] + (1-\alpha)[m(B) - m(A)]\\ &= \alpha m + (1-\alpha)m'(B) - \alpha m + (1-\alpha)m'(A)\end{aligned}$$

Let $(A_n)_n$ be a countable mutually disjoint sequence of sets in $\mathscr{E}$.

$$\begin{aligned}(\alpha m + (1-\alpha)m')(\cup_n A_n) &= \alpha m(\cup_n A_n) + (1-\alpha)m'(\cup_n A_n)\\ &= \alpha \sum_n m(A_n) + (1-\alpha)\sum_n m'(A_n)\\ &= \sum_n \alpha m(A_n) + (1-\alpha)m'(A_n)\end{aligned}$$

Note that we used the sigma-additivity to multiply and regroup each series.
We can now prove that the potential function is convex. Let m and m' be two probability measures. $\alpha \in$ [0;1]. We adopt the same notations.

$$\begin{aligned}F(\alpha m + (1-\alpha)m') &= \int_{\mathbb{R}^d} \Phi(y - \mathbb{E}^{\alpha m + (1-\alpha)m'}[\overline{\phi}(X,z)])\nu(dy,dz)\\ &= \int_{\mathbb{R}^d} \Phi(y - \alpha\mathbb{E}^{m}[\overline{\phi}(X,z)] + (1-\alpha)\mathbb{E}^{m'}[\overline{\phi}(X,z)])\nu(dy,dz)\\ &= \int_{\mathbb{R}^d} \Phi(\alpha y + (1-\alpha)y - \alpha\mathbb{E}^{m}[\overline{\phi}(X,z)] + (1-\alpha)\mathbb{E}^{m'}[\overline{\phi}(X,z)])\nu(dy,dz)\\ &\overset{\Phi convex}{\leq} \int_{\mathbb{R}^d} \alpha\Phi(y - \mathbb{E}^{m}[\overline{\phi}(X,z)]) + (1-\alpha)\Phi(y - \mathbb{E}^{m'}[\overline{\phi}(X,z)])\nu(dy,dz)\\ &\leq \alpha\int_{\mathbb{R}^d} \Phi(y - \mathbb{E}^{m}[\overline{\phi}(X,z)])\nu(dy,dz) + (1-\alpha)\int_{\mathbb{R}^d} \Phi(y - \mathbb{E}^{m'}[\overline{\phi}(X,z)])\nu(dy,dz)\end{aligned}$$

The last inequality holding from linearity of the integral. □

One of the main points is the characterization of the first order condition and the minimizer(s) of a convex function F : $\mathscr{P}(\mathbb{R}^d) \to \mathbb{R}^d$. We'll now comment the pertinency of each assumption and remark. Should the need arise we'll try to detail proofs.
Assumption 2.1: Choosing $F \in C^1$ and bounded from below gives well-posedness of the problem.
Assumption 2.2: We obtain that the 'Frechet' differential of U(x) at point x is bounded from below with the Lispchitz continuity we deduce the inequality.

Characterization of the minimizer Proposition 2.4: according to the authors the main goal of their article is to characterize the minimizer of V^σ. Before we get to explaining the proof provided by the authors we notice that $H = \mathbb{E}^m[log(m/g)]$
To find the minimizer Γ-convergence is used. The idea is that in the weak topology the sequence of minimizers will converge towards the minimizer. Proof is given in proposition 2.3. This proposition is crucial as they are able to extract a sequence of minimizers that converges to the minimizer of the free Energy function. A quick comment about the proof of said proposition page 12, symbol * denotes convolution.

2.1.1 Existence of a solution

The first step is to establish the existence of a solution to the minimization problem. As the problem is not degenerate V is finite for at least a probability. The authors then define:

$$S = \{m : \frac{\sigma^2}{2} H(m) \leq V^\sigma(\overline{m}) - inf_{m' \in P(\mathbb{R}^d)} F(m')\}$$

"*We recall that such relative entropy H has the properties: it is strictly convex when restricted to measures absolutely continuous with g, it is weakly lower semi-continuous and its sub-level sets are compact.*". Our first point of interest is compactnesss of the sub-level sets. It follows from Prohorov's theorem and Donsker-Varadhan variational formula. Detail proof of the Donsker-Varadhan formula is given in [20] section 1.4. We also use that F is convex then continuous then lower semi-continuous. V is lower semi-continuous as finite sum of lower semi-continuous functions. Minimum of lower-semi continuous function on compact set is attained. Still is it the global minimum?
$\forall m \notin S, V^\sigma(m) \geq V^\sigma(\overline{m})$. Therefore the minimum on S is the global minimum.

2.1.2 Unicity of the solution

We'd like the problem to have a unique solution. To prove this the researchers base themselves on the following proposition

Proposition 2.2. *Let F: $X \to Y$ a strictly convex function, the minimization problem $min_{x \in X} F(x)$ admits a unique solution, should there be one.*

Moreover denoting m^* the solution to the problem we have, m $\in S \implies H(m*) < \infty$. "*Therefore, m is absolutely continuous with respect to the Gibbs measure, so also absolutely continuous with respect to the Lebesgue measure*" This true as absolute continuity is a transitive relation.

2.1.3 Sufficient condition

Define $I_\sigma = \{m \in P(\mathbb{R}^d) : \frac{\delta F}{\delta m}(m^*, \cdot) + \frac{\sigma^2}{2} log(m) + \frac{\sigma^2}{2} U$ is a constant}. m^* is equivalent to the Lebesgue measure thanks to absolute continuity. Choosing another measure m equivalent to Lebesgue's they define its density with respect to the minimizer. Defining a mixture of measures m^ε they study the lower bounds of F and H as both are lower semi-continuous. Using lemma 4.1 and that $\frac{\delta F}{\delta m}$ is bounded continuous, we deduce the first inequality. To obtain the second inequality one uses log properties and the identity: $m^\varepsilon - m^* = (f-1)m^*$ with the expression of g as an exponential function.

2.1.4 Necessary condition

Assuming m^* is a minimizer of V^σ and by computing the same "limit" quantity as before. Using dominated convergence theorem with respect to ε the necesary condition is proved. Usually we'd have an undetermined form tending ε towards 0 to avoid this we use the mean inequality(Integral form).

3 Langevin Mean-Field Dynamics

One way to compute the minimum of the Free-energy function is to use the Langevin Mean-Field Dynamics that approximates the minimizer of the free energy function (V$^\sigma$) by using the marginal laws of the following mean-field Langevin equation :

$$\mathrm{d}X_t = -\left(D_m F\left(m_t, X_t\right) + \frac{\sigma^2}{2} \nabla U\left(X_t\right)\right) \mathrm{d}t + \sigma \mathrm{d}W_t$$

where m_t is the law of X_t and $(W_t)_{t \geq 0}$ is a standard d-dimensional Brownian Motion and $\sigma \in \mathbb{R}^+$
Remark that a Langevin equation can be solved either by Monte-Carlo simulation or by Fokker-Planck equation. In the paper, authors have chosen this last method.

3.1 The Fokker-Planck equation

Under some Assumptions on U as U$\in C^\infty$ (see Assumption 2.2 of the paper for more details) and on $D_m F$ as bounded and Lipschitz continuous (see Assumption 2.6)
Assume also $m_0 \in \mathscr{P}_p(\mathbb{R}^d)$ for some $p \geq 2$.
We have the below intrinsic Fokker-Planck equation:

$$\partial_t m = \nabla \cdot \left(b(x, m)m + \frac{\sigma^2}{2} \nabla m\right)$$

where m is the law of Langevin equation solution and $b(x, m) = \left(D_m F\left(m, x\right) + \frac{\sigma^2}{2} \nabla U\left(x\right)\right)$
m is also the unique solution to Fokker-Planck equation such that $t \to m_t$ is weakly continuous on $[0, \infty)$ and the joint density function of $m : (t, x) \to h(t, x)$ exists and $h \in C^{1,\infty}((0, \infty) \times \mathbb{R}^d, \mathbb{R})$
Let notice that m can be viewed as the probability of a particle at time t in the $\mathbb{R}^d$-space.

3.2 Existence of an invariant measure, minimizer of the free energy function

In this section we'll interest in the convergence of $(m_t)_{t \in \mathbb{R}^+}$, in particular under the same assumptions that 2.1, and assumptions on F that is convex, bounded,... (see Assumption 2.1 of the paper for details) and $m_0 \in \bigcup_{p>2} \mathscr{P}_p(\mathbb{R}^d)$:

There exists an invariant measure of Langevin Equation equal to $m^* := argmin_m V^\sigma(m)$ and $(m_t)_{t \in \mathbb{R}^+}$ converges to m^*

This above result is the key theorem of the paper since it links Langevin Dynamics and the minimizer of the free energy function.
We'll make a sketch step by step of the proof of this important theorem.

3.2.1 Sketch of the Proof

Step 1 :
Firstly, we define a dynamic system $S(t)[m_0] := m_t$ and a w-limit set such that:

$$w\left(m_0\right) := \left\{\mu \in \mathscr{P}_2\left(\mathbb{R}^d\right) : \text{ there exist } t_n \rightarrow \infty \text{ such that } \mathscr{W}_2\left(S\left(t_n\right)\left[m_0\right], \mu\right) \rightarrow 0\right\}$$

We know that the above set is nonempty, compact and invariant. To prove this last result we use that $S(t)$ is continuous with the $\mathscr{W}_2$-topology, results on compact set and some results for our specific problem (refer you Proposition 2.6 in the paper for more details).
As $w(m_0)$ is compact (bounded and closed with $\mathscr{W}_2$-topology), there exists $\tilde{m} \in argmin_{m \in w(m_0)} V^\sigma(m)$.
By the invariance of $w(m_0)$, for $t > 0$ there exists a probability measure $\mu \in w(m_0)$ such that $S(t)[\mu] = \tilde{m}$
Under the same assumption that this section, we have for any t > s > 0

$$V^\sigma\left(m_t\right) - V^\sigma\left(m_s\right) = -\int_s^t \int_{\mathbb{R}^d}\left|D_m F\left(m_r, x\right) + \frac{\sigma^2}{2} \frac{\nabla m_r}{m_r}(x) + \frac{\sigma^2}{2} \nabla U(x)\right|^2 m_r(x) \mathrm{d} x \mathrm{d} r \ (1)$$

This equality is the result of the Itô formula applied to F,H and so V, and some relations (see Lemma 6.4 in the paper for relations).
Due to the non-negative norm and the probability measure we have for any s > 0 :

$$V^\sigma(S(t+s)[\mu]) \leq V^\sigma(\tilde{m})$$

Since $w(m_0)$ is invariant, $S(t+s)[\mu] \in w(m_0)$ and so $V^\sigma(S(t+s)[\mu]) = V^\sigma(\tilde{m})$
With (1) we obtain:

$$0 = \frac{dV^\sigma(S(t)[\mu])}{dt} = -\int_{\mathbb{R}^d}\left|D_m F(\tilde{m}, x) + \frac{\sigma^2}{2} \frac{\nabla \tilde{m}}{\tilde{m}}(x) + \frac{\sigma^2}{2} \nabla U(x)\right|^2 \tilde{m}(x) \mathrm{d} x$$

$\tilde{m} = S(t)[\mu]$ is equivalent to the Lebesgue measure that results from $i)$ Lemma 6.1 in the paper where the solution of Langevin Equation is equivalent to the scaled Wiener Measure. Finally we obtain:

$$D_m F(\tilde{m}, \cdot) + \frac{\sigma^2}{2} \frac{\nabla \tilde{m}}{\tilde{m}} + \frac{\sigma^2}{2} \nabla U = 0 \ (2)$$

The probability measure $\tilde{m}$ is an invariant measure of Langevin equation since it's a stationary solution to the Fokker-Planck equation.
Due to the first order condition, $\tilde{m} = m^*$

Step 2 :
Let $(m_{t_n})_n$ be the subsequence converging to m^*. This step aims at proving that $V^\sigma(m^*) = lim_{n \to \infty} V^\sigma(m_{t_n})$
As F is continuous, it's sufficient to prove $\int_{\mathbb{R}^d} m^* \log\left(m^*\right) \mathrm{d}x = \lim_{n\to\infty} \int_{\mathbb{R}^d} m_{t_n} \log\left(m_{t_n}\right) \mathrm{d}x$. Since the entropy is lower-semicontinuous, it's enough to prove that

$$\int_{\mathbb{R}^d} m^* \log\left(m^*\right) \mathrm{d} x \geq \varlimsup_{n \rightarrow \infty} \int_{\mathbb{R}^d} m_{t_n} \log\left(m_{t_n}\right) \mathrm{d} x$$

By (2) $-log(m^*)$ is semi-convex, thus with the HWI inequality (it links the relative entropy H and the relative Fisher information I_n), we obtain:

$$\int_{\mathbb{R}^d} m_{t_n}\left(\log\left(m_{t_n}\right) - \log\left(m^*\right)\right) \mathrm{d} x \leq \mathscr{W}_2\left(m_{t_n}, m^*\right)\left(\sqrt{I_n} + C \mathscr{W}_2\left(m_{t_n}, m^*\right)\right)$$

with :

$$I_n := \mathbb{E}\left[\left|\nabla \log\left(m_{t_n}\left(X_{t_n}\right)\right) + \frac{2}{\sigma^2} D_m F\left(m^*, X_{t_n}\right) + \nabla U\left(X_{t_n}\right)\right|^2\right]$$

We'll prove that $sup_n I_n < \infty$. Since $D_m F$ bounded and is of linear growth, by Lemma 5.1 (where we obtain some results on convergence refer you to Lemma 5.1 for more details), we have:

$$\sup_n \mathbb{E}\left[\left|\frac{2}{\sigma^2} D_m F\left(m^*, X_{t_n}\right) + \nabla U\left(X_{t_n}\right)\right|^2\right] < \infty$$

Then as ∇b is bounded and with $ii)$ Lemma 6.2 and thanks to the result on the exponential martingale that is conditionally $\mathbb{L}^2$-differentiable we obtain:

$$\mathbb{E}\left[\left|\nabla \log\left(m_{t_n}\left(X_{t_n}\right)\right)\right|^2\right] < \infty$$

So $sup_n I_n < \infty$ and the HWI inequality becomes:

$$\int_{\mathbb{R}^d} m_{t_n}\left(\log\left(m_{t_n}\right) - \log\left(m^*\right)\right) \mathrm{d} x \leq C \mathscr{W}_2\left(m_{t_n}, m^*\right)\left(1 + \mathscr{W}_2\left(m_{t_n}, m^*\right)\right)$$

Finally as $n \to \infty$, $\mathscr{W}_2\left(m_{t_n}, m^*\right) \to 0$ and we have :

$$V^\sigma(m^*) = lim_{n \to \infty} V^\sigma(m_{t_n})$$

Step 3 :
In this section we'll prove the convergence of $(m_t)_{t \in \mathbb{R}^+}$. As $V^\sigma(m_t)$ as non-increasing in t and bounded below. There exists a constant $c := lim_{t \to \infty} V^\sigma(m_t)$, by monotone convergence theorem.
As a subsequence of a convergent sequence converges to the same limit, by the Step 2, $c = V^\sigma(m^*)$
By Step 1, for any $\mu \in w(m_0)$, there exists a subsequence $(m_{t_n})_n$ converging to $\mu \in w(m_0)$ and, by lower-semicontinuity, we have $V^\sigma(\mu) \leq \lim inf_{n \to \infty} V\left(m_{t_n}\right) = c$. Since $m^* = \tilde{m} = argmin_{m \in w(m_0)} V^\sigma(m)$. We obtain for all $\mu \in w\left(m_0\right)$:

$$V^\sigma(\mu) = V^\sigma\left(m^*\right) = c$$

So $w\left(m_0\right) = \{m^*\}$, which is, $\lim_{t \to \infty} \mathscr{W}_2\left(m_t, m^*\right) = 0$

4 Convergence speed of the flow of marginal laws

In this section we'll discuss about the convergence speed of the marginal laws that is exponential.
Under suitable assumptions on D_mF, U and constants ie D_mF is Lipschitz and Linear Growth, ∇U is Linear Growth and some growth conditions on differential (see Assumption 2.11 for more details). We obtain with F convex,bounded and $F\in\mathscr{C}^1$ the following theorem:

$$\mathscr{W}_2\left(m_t,m^*\right)\leq e^{(6C_F-C_U)t}\mathscr{W}_2\left(m_0,m^*\right)$$

with $(m_t)_{t\in\mathbb{R}+}$, the flow of marginal laws of Langevin Equation solution.
As $6C_F-C_U\leq 0$ (due to $C_F<<C_U$), the flow converges very rapidly towards the stationary measure.
Let notice that the conditions on constants σ and U become more restrictive than Assumption 2.2, on the contrary, conditions on D_mF allow more candidate for F than Assumption 2.6.

4.1 Proof

Now, we'll proof the exponential convergence.
As 5.1.1, there is $m^*\in\mathscr{P}_2\left(\mathbb{R}^d\right)$, an invariant measure of the Langevin equation. Let $(X_t')_{t\geq 0}$ denote the solution of Langevin equation starting from $X_0'\sim m^*$, then $X_t'\sim m^*$ for all $t\geq 0$ and $(X_t)_{t\geq 0}$ another solution we have for all $t\geq 0$ if $\sup_{t>0}\mathbb{E}\left|X_t\right|^p<\infty$ for $p>2$:

$$\mathbb{E}\left[\left|X_t-X_t'\right|^2\right]\leq e^{(6C_F-C_U)t}\mathbb{E}\left[\left|X_0-X_0'\right|^2\right]$$

This inequality is obtained by applying result from Lyapunov function, for more details see Lemma 7.1.
Finally, by simultaneously taking infimum over all coupling of m_t and m^* and over all coupling of m_0 and m^* we obtain the exponential convergence.

5 Advantages and links to the course

In this section, we shall define the problem of minimization from the main results to a neural network. Next, we will see that with 1-hidden layer there is convergence of the marginal laws of the corresponding mean-field Langevin dynamics to the optimal weight of the neural network. After, we will elaborate on the cases of fully connected deep neural network.

Define $\varphi:\mathbb{R}\to\mathbb{R}$ a locally Lipschitz function, $l\in\mathbb{N}$ and define $\varphi^l:\mathbb{R}^l\to\mathbb{R}^l$, for $z=(z_1,\ldots,z_l)^\top$, we have $\varphi^l(z)=\left(\varphi\left(z_1\right),\ldots,\varphi\left(z_1\right)\right)^\top$. Fix $L\in\mathbb{N}$ the number of layers, $l_k\in\mathbb{N}, k=0,1,\ldots L-1$ the size of input to layer k and $l_L\in\mathbb{N}$ the size of the network output. Define a fully connected artificial network $\Psi=\left(\left(\alpha^1,\beta^1\right),\ldots,\left(\alpha^L,\beta^L\right)\right)\in\Pi$, where for $k=1,\ldots,L$, we have real $l^k\times l^{k-1}$ matrices α^k and real l^k-dimensional vectors β^k. We deduce that $\Pi=\left(\mathbb{R}^{l^1\times l^0}\times\mathbb{R}^{l^1}\right)\times\left(\mathbb{R}^{l^2\times l^1}\times\mathbb{R}^{l^2}\right)\times\cdots\times\left(\mathbb{R}^{l^L\times l^{L-1}}\times\mathbb{R}^{l^L}\right)$. The artificial neural network gives a reconstruction function $\mathscr{R}\Psi:\mathbb{R}^{l^0}\to\mathbb{R}^{l^L}$ recursively, for $z_0\in\mathbb{R}^0$:

$$\left(\mathscr{R}\Psi\right)\left(z^0\right)=\alpha^Lz^{L-1}+\beta^L, z^k=\varphi^{l^k}\left(\alpha^kz^{k-1}+\beta^k\right), k=1,\ldots,L-1$$

Defining α_k^i, β_i^k the i-th row of the matrix α^k and β^k, equivalently we define the reconstrution funtion :

$$\left(\mathscr{R}\Psi\right)\left(z^0\right)_i=\alpha_i^L\cdot z^{L-1}+\beta_i^L,\quad \left(z^k\right)_i=\varphi\left(\alpha_i^k\cdot z^{k-1}+\beta_i^k\right), k=1,\ldots,L-1$$

We need to find the parameters Ψ in which the artificial neural network provides a good approximation to a real world problem in supervised learning. Defining the potential function ϕ, and having training data $(y^j,z^j)_{j=1}^N, (y_j,z_j)\in\mathbb{R}^d$, we can approximate the optimal parameters by solving :

$$\underset{\Psi\in\Pi}{\operatorname{argmin}}\tfrac{1}{N}\textstyle\sum_{j=1}^N\Phi\left(y^j-\left(\mathscr{R}\Psi\right)\left(z^j\right)\right)$$

Having huge data sets permits us to use the law of large numbers, and if that training data is distributed to some measure ν with compact support, we can rewrite the problem as :

$$\underset{\Psi\in\Pi}{\operatorname{argmin}}\textstyle\int_{\mathbb{R}^d}\Phi(y-(\mathscr{R}\Psi)(z))\nu(\mathrm{d}y,\mathrm{d}z)$$

Which gives us a non-convex minimization problem, and goes on par with our theoretical problem.
The universal approximation theorem is applied here with $(\mathscr{R}\Psi)$ in the second definition, where the activation functionφ mentioned to be bounded, continuous and non-constant.

5.1 Fully connected 1-hidden layer neural network

With L=2, take $d\in\mathbb{N}$, $n\in\mathbb{N}$. We will consider the 1-hidden layer neural network for approximating functions from $\mathbb{R}^d$ to $\mathbb{R}$ where: $l_0=d$, $l_1=n$, $\beta^2=0\in\mathbb{R}$, $\beta^1=0\in\mathbb{R}^n$, $\alpha^1\in\mathbb{R}^{n\times d}$. Take $\alpha^2=(\frac{c_1}{n},\ldots,\frac{c_n}{n})^\top$, and the $c_i\in\mathbb{R}$. We deduce the neural network $\Psi^n=\left(\left(\alpha^1,\beta^1\right),\left(\alpha^2,\beta^2\right)\right)$. For $z\in\mathbb{R}^{l^0}$, the reconstruction function is :

$$\left(\mathscr{R}\Psi^n\right)(z)=\alpha^2\varphi^{l^1}\left(\alpha^1z\right)=\tfrac{1}{n}\textstyle\sum_{i=1}^nc_i\varphi\left(\alpha_i^1\cdot z\right)$$

With law of large numbers, and under some assumptions : $\frac{1}{n}\sum_{j=1}^nc_j\varphi\left(\alpha_j^1\cdot z\right)\to\mathbb{E}^m[B\varphi(A\cdot z)]$ as $n\to\infty$ with m the law of (B,A) and $\mathbb{E}^m$ the expectation under m. We get the minimization problem :

$$\min_{m\in\mathscr{P}\left(\mathbb{R}^d\times\mathbb{R}\right)}F(m),\qquad\text{where}\qquad F(m):=\textstyle\int_{\mathbb{R}^d}\Phi\left(y-\mathbb{E}^m[l(B)\varphi(A\cdot z)]\right)\nu(\mathrm{d}z,\mathrm{d}y)$$

with $l:\mathbb{R}\to K$, K some compact set, a truncation function.
If Φ is convex, the objective function F becomes a convex function on $\mathscr{P}(\mathbb{R}^d)$.

5.2 Deep neural network

In the 1-hidden layer case, we have seen the linearization of the problem. We are going to see that there are case in which this linearization technique applies, as there is that does not.

Averaging fully connected deep artificial neural networks

Take n fully connected artificial networks with L hidden layers and :

$$\Psi^{(i)} = \left(\left(\alpha^{(i,1)}, \beta^{(i,1)}\right), \left(\alpha^{(i,2)}, \beta^{(i,2)}\right), \ldots, \left(\alpha^{(i,L)}, \beta^{(i,L)}\right)\right) \in \Pi, i = 1, \ldots, n$$

We construct the artificial neural network from the previous n networks. And let $\Psi^n := \left(\left(\alpha^1, \beta^1\right), \left(\alpha^2, \beta^2\right), \ldots, \left(\alpha^L, \beta^L\right)\right) \in \Pi^n$. For $z^0 \in \mathbb{R}^{l^0}$, we have :

$$\left(\mathscr{R}\Psi^n\right)\left(z^0\right) = \tfrac{1}{n}\textstyle\sum_{i=1}^n \left(\mathscr{R}\Psi^{(i)}\right)\left(z^0\right)$$

Take $m^n = \frac{1}{n}\sum_{i=1}^n \delta_{\left(\alpha^{(i,1)}, \beta^{(i,1)}\right), \left(\alpha^{(i,2)}, \beta^{(i,2)}\right), \ldots, \left(\alpha^{(i,L)}, \beta^{(i,L)}\right)} \in \mathscr{P}(\Pi)$ the empirical measure over the parameter space fully describing the network Ψ^n. We can rewrite :

$$\left(\mathscr{R}\Psi^n\right)(z) = \textstyle\int_\Pi (\mathscr{R}x)(z) m^n(\mathrm{d}x)$$

We remark that $m^n \mapsto \int_\Pi (\mathscr{R}x)(z) m^n(\mathrm{d}x)$ is a linear then convex function on the measure m^n.

Fully connected 2-hidden layers neural network

This case will show that the linearization technique can't be applied to all fully connected deep artificial neural networks.

L=3 and take $\beta^3 = \beta^2 = \beta^1 = 0$. Let $\alpha^3 = \left(\frac{c_1^3}{l^3}, \cdots, \frac{c_{l^3}^3}{l^3}\right)^\top$, with $c_i^3 \in \mathbb{R}$, $\alpha_{ij}^2 := \frac{c_{ij}^2}{l^2}$. The neural network is given by $\Psi^{l^3, l^2} = \left(\left(\alpha^1, \beta^1\right), \left(\alpha^2, \beta^2\right), \left(\alpha^3, \beta^3\right)\right)$.

The reconstruction function is :

$$\left(\mathscr{R}\Psi^{l^3, l^2}\right)(z) = \alpha^3 \varphi^{l_2}\left(\alpha^2 \varphi^{l_1}\left(\alpha^1 z\right)\right) = \tfrac{1}{l^3}\textstyle\sum_{i=1}^{l^3} c_i^3 \varphi\left(\tfrac{1}{l^2}\textstyle\sum_{j=1}^{l^2} c_{ij}^2 \varphi\left(\alpha_j^1 z\right)\right)$$

Taking an empirical law $\mu^{I, l^2} := \frac{1}{l^2}\sum_{j=1}^{l^2} \delta_{\left\{\alpha^2, \alpha^1\right\}}$, conditioning on I an uniformly distributed random variable on the support $\left\{c_1^3, c_2^3, \ldots, c_{l^3}^3\right\}$. We can rewrite :

$$\left(\mathscr{R}\Psi^{l^3, l^2}\right)(z) = \mathbb{E}^I\left[c^I \varphi\left(\textstyle\int_{\mathbb{R}\times\mathbb{R}^{l_0}} y^2 \varphi\left(y^1 z\right) \mu^{I, l_1}\left(\mathrm{d}y^1, \mathrm{d}y^2\right)\right)\right]$$

Which is not in general a convex function anymore.

5.3 The minimizer and the parameters

In the static case, we have the statement from theorem 8.1 :

If there is an $m^* \in \mathscr{P}_2\left(\mathbb{R}^d\right)$ such that $F\left(m^*\right) = \inf_{m \in \mathscr{P}_2\left(\mathbb{R}^d\right)} F(m)$ then with i.i.d $\left(X_i^*\right)_{i=1}^N$ such that $X_i^* \sim m^*$ $i = 1, \ldots, N$ we have that

$$\left|\mathbb{E}\left[F\left(\frac{1}{N}\sum_{i=1}^N \delta_{X_i}\right)\right] - F\left(m^*\right)\right| \leq \frac{2L}{N} \text{ and } \left|\inf_{\left(x_i\right)_{i=1}^N \subset \mathbb{R}^d} F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right) - F\left(m^*\right)\right| \leq \frac{2L}{N}$$

It shows that the minimizer of the free energy function can be found by taking the amount of layers to infinity. It will be proved by using the properties of linear functional derivatives and measures.

In the dynamic case, we get the classical Langevin dynamics on $(\mathbb{R}^d)^N$:

$$\mathrm{d}X_t^i = \left(\textstyle\int_{\mathbb{R}^d} \dot{\Phi}\left(y - \tfrac{1}{N}\textstyle\sum_{j=1}^N \hat{\varphi}\left(X_t^j, z\right)\right) \nabla\hat{\varphi}\left(X_t^i, z\right) \nu(\mathrm{d}z, \mathrm{d}y) - \tfrac{\sigma^2}{2}\nabla U\left(X_t^i\right)\right) \mathrm{d}t + \sigma \mathrm{d}W_t^i$$

where $\forall z \in \mathbb{R}^{d-1} : \nabla\hat{\varphi}\left(x^i, z\right) = \nabla_{\left(\beta^i, \alpha^i\right)}\left[\ell\left(\beta^i\right)\varphi\left(\alpha^i z\right)\right] = \begin{pmatrix} \dot{\ell}\left(\beta^i\right)\varphi\left(\alpha^i \cdot z\right) \\ \ell\left(\beta^i\right)\dot{\varphi}\left(\alpha^i \cdot z\right) z \end{pmatrix}$

Considering the implementable algorithm with time discretisation, and a time step τ, the Euler scheme gives an explicit result. And if we have data points iid samples from ν, taking the loss function Φ to be the square loss function, a version of the gradient descent algorithm for the evolution of the parameter x_k^i will be given by :

$$x_{k+1}^i = x_k^i + 2\tau\left(\left(y_k - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}\left(x_k^j, z_k\right)\right)\nabla\hat{\varphi}\left(x_k^i, z^k\right) - \frac{\sigma^2}{2}\nabla U\left(x_k^i\right)\right) + \sigma\sqrt{\tau}\xi_k^i$$

with ξ_k^i being independent samples of $\mathscr{N}(0, I_d)$.

Let us note that the stochastic gradient descent only slightly deviates from the usual gradient descent. Indeed, instead of going through all the gradients of the different datas, it picks at iteration k, a random variable i_k uniformly distributed on $\{1, ..., N\}$. This deviation gives this algorithm a higher speed but is unstable in the sense where if we reiterate the algorithm many times, it's possible that we will not find the same results. In the course, at iteration k : $\theta^{k+1} \leftarrow \theta^k - \eta_k \nabla E_{i_k}(\theta^k)$.

We can find a deterministic part and a random part in x_{k+1}^i, which explains the link between the evolution of x_i and the stochastic gradient descent.

5.4 Numerical application : Averaging Deep Artificial Neural Networks

As presented before, the averaging deep neural networks corresponds to our theoretical study. In this section, we will study the performances between the averaging deep artificial networks and the fully connected ones. In order to do that, we will approximate partial differential equations.

Consider :

$$\begin{cases} \partial_t v + \operatorname{tr}\left(a\partial_x^2 v\right) + b\partial_x v = 0 \quad \text{in } [0, T) \times D \\ v(T, \cdot) = g \text{ on } D \end{cases}$$

with $a := \frac{1}{2}\sigma\sigma^*$, $v \in C^{1,2}([O, T] \times D)$. Take the SDE :

$$\mathrm{d}X_s = b\left(X_s\right)\mathrm{d}s + \sigma\left(X_s\right)\mathrm{d}W_s t \in [t, T], X_t = x$$

From Feynman-Kac formula, with a Markov process X solution of this SDE, we have :

$$v(t,x) := \mathbb{E}[g(X_T) \mid X_t = x]$$

Taking a partition of $[0,T] : \pi := \{t = t_0 < ... < t_{N_{steps}} = T\}$
Consider ΔW_{t_k} the increment at k, $k = 1, ..., N_{steps}$, $W^{\pi} := (W_{t_i})_{i=1}^{N_{steps}}$. Define $(X^{\pi}_{t_i})_{i=1}^{N_{steps}}$ an approximation from the SDE, and :

$$X^{\pi}_{t_{k+1}} = G(X^{\pi}_{t_k}, \Delta W_{t_{k+1}})$$

where $G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ a measurable function, andin the Euler scheme case, we have : $G(x,y) = x + b(x)h + \sigma(x)y$.
Under assumptions, the fact that the sum and composition of deep neural networks is still a deep neural network, we construct $\Psi^{(W^{\pi})}$ and get $X^{\pi}_T = (\mathscr{R}\Psi^{(W^{\pi})}(X_0)$. Assume next that there is a network $\Psi^{(g)}$ such that $\forall x \in \mathbb{R}^d$, we have $g(x) = (\mathscr{R}\Psi^{g}(x)$. Approximating $v(t,x)$ gives :

$$v(t,x) = \mathbb{E}\left[g\left(X_T\right)\right] \approx \mathbb{E}\left[g\left(X^{\pi}_T\right)\right] = \mathbb{E}\left[\left(\mathscr{R}\Psi^{(g)}\right)\left(X^{\pi}_T\right)\right] = \mathbb{E}\left[\left(\mathscr{R}\Psi^{(g)}\right) \circ \left(\mathscr{R}\Psi^{(W^{\pi})}\right)(x)\right]$$

Using n iid samples $(W^{\pi,(j)})_{j=1}^{n}$ from W^{π}, we have similarity with our averaging formula :

$$v(t,x) = \tfrac{1}{n}\textstyle\sum_{j=1}^{n}\left(\mathscr{R}\Psi^{(g)}\right) \circ \left(\mathscr{R}\Psi^{\left(W^{\pi,(j)}\right)}\right)(x)$$

In the numerical part, we can observe that the averaging artificial neural networks makes a good approximation although less precise than the longer trained and larger fully connected neural network.

6 Relevance of the article

6.1 Context and breakthrough of the article

Neural networks have achieved immense practical success over the past decade. Neural networks are nonlinear statistical models whose parameters are estimated from data using gradient descent. They have been employed as critical components of many important technologies in a variety of industries. This practical success has sparked significant interest in their mathematical analysis. Currently, there is limited mathematical understanding of neural networks, but there is overwhelming empirical evidence that deep neural networks trained with stochastic gradient descent perform (extremely) well in high dimensional setting. Nonetheless, complete mathematical theory that would provide theoretical guarantees why and when these methods work so well has been elusive.

In a recent series of works, the task of learning the optimal weights in deep neural networks is viewed as a sampling problem. The picture that emerges is that the aim of the learning algorithm is to find optimal distribution over the parameter space (rather than optimal values of the parameters). As a consequence, individual values of the parameters are not important in the sense that different sets of weights sampled from the correct (optimal) distribution are equally good. To learn optimal weights, one needs to find an algorithm that samples from the correct distribution.

In “Mean-field Langevin dynamics and energy landscape of neural networks” paper, it has been shown that in the case of one-hidden layer network and in the case of ensembles of deep neural networks the stochastic gradient algorithm does precisely that. The key mathematical tools to these results turn out to be differential calculus on the measure space.

Indeed, the work performed by Kaitong Hu, Zhenjie Ren, David Siska and Lukasz Szpruch aims at providing a theoretical underpinning for the convergence of stochastic gradient type algorithms widely used for non-convex learning tasks such as training of deep neural networks. The key insight is that the certain class of finite dimensional non-convex problems becomes convex when lifted to infinite dimensional space of measures. This observation is used to show that the corresponding energy functional defined on the space of probability measures has a unique minimizer which can be characterized by a 1rst order condition using the notion of linear functional derivative. It is proved then that the flow of marginal laws induced by the mean-field Langevin dynamics converges to the stationary distribution which is exactly the minimizer of the energy functional. The authors also show that this convergence is exponential under mild conditions. The analysis is based on a pathwise perspective on Otto calculus. This proof of convergence to stationary probability measure is novel and it relies on a generalization of LaSalle’s invariance principle.

Other groups developed similar mean-field description of non-convex learning problems. In particular the pioneering works of “A mean field view of the landscape of two-layer neural networks”, “On the global convergence of gradient descent for over-parameterized models using optimal Transport” as well as “Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error” proved convergence of gradient algorithms to the minimum using the theory of gradient flow in the Wasserstein space of probability distributions. Those articles will be presented briefly bellow.

Results in “A mean field view of the landscape of two-layer neural networks” are the closest to this paper but the proofs are different. It seems that main differences are that the “Mean-field Langevin dynamics and energy landscape of neural networks” paper provide a probabilistic perspective, generalize and provide complete proofs of some key results such as chain rule for the flows of measures (Theorem 2.8) and global convergence of flow of measures to the invariant measure (Theorem 2.10). In particular it establishes convergence to the invariant measure in 2-Wasserstein distance and also demonstrates that for sufficiently regularized problem that convergence is exponential. Furthermore the paper deals with general loss function.

6.2 Summary of related articles

A mean field view of the landscape of two-layer neural networks - 14/08/2018

Song Mei, Theodor Misiakiewicz, Andrea Montanari

Learning a neural network requires optimizing a nonconvex high-dimensional objective (risk function), a problem that is usually attacked using stochastic gradient descent (SGD). Does SGD converge to a global optimum of the risk or only to a local optimum? In the former case, does this happen because local minima are absent or because SGD somehow avoids them? In the latter, why do local minima reached by SGD have good generalization properties? In this paper, the authors study the case of two-layer networks and derive a compact description of the SGD dynamics in terms of a limiting partial differential equation. They prove that, in a suitable scaling limit, SGD dynamics is captured by a certain nonlinear partial differential equation (PDE) that they call distributional dynamics. Considering several specific examples they show how distributional dynamics can be used to prove convergence of SGD to networks with nearly ideal generalization error.

Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit - 19/02/2019

Song Mei, Theodor Misiakiewicz, Andrea Montanari

The authors consider learning two layer neural networks using stochastic gradient descent. The mean-field description of this learning dynamics approximates the evolution of the network weights by an evolution in the space of probability distributions. This evolution can be defined through a partial differential equation or, equivalently, as the gradient flow in the Wasserstein space of probability distributions. In this paper, it is established stronger and more general approximation guarantees for the mean field description than earlier works (number of hidden units only needs to be larger than a quantity dependent on the regularity properties of the data, and independent of the dimensions, generalization to the case of unbounded activation functions, extention to noisy stochastic gradient descent).

Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error - 22/05/2018
Grant M. Rotskoff, Eric Vanden-Eijnden
There are few rigorous results about the representation error and trainability of neural networks. The authors endeavor to characterize both the error and the scaling of the error with the size of the network by reinterpreting the standard optimization algorithm used in machine learning applications, stochastic gradient descent, as the evolution of a particle system with interactions governed by a potential related to the objective or "loss" function used to train the network. They show that, when the number n of parameters is large, the empirical distribution of the particles descends on a convex landscape towards a minimizer at a rate independent of n. They establish a Law of Large Numbers and a Central Limit Theorem for the empirical distribution, which together show that the approximation error of the network universally scales as O$(n-1)$.

Analysis of a two-layer neural network via displacement convexity - 05/01/2019
Adel Javanmard, Marco Mondelli, Andrea Montanari
Fitting a function by using linear combinations of a large number N of 'simple' components is one of the most fruitful ideas in statistical learning. Unfortunately, little is known about global convergence properties of the resulting non-convex (in general) risk minimization problem solved by gradient descent or its variants. Here, the authors consider the problem of learning a concave function f on a compact convex domain $\Omega \subset \mathbb{R}^d$, using linear combinations of 'bump-like' components (neurons). The parameters to be fitted are the centers of N bumps, and the resulting empirical risk minimization problem is highly non-convex. They prove that, in the limit in which the number of neurons diverges, the evolution of gradient descent converges to a Wasserstein gradient flow in the space of probability distributions over Ω. Further, when the bump width δ tends to 0, this gradient flow has a limit which is a viscous porous medium equation. Remarkably, the cost function optimized by this gradient flow exhibits a special property known as displacement convexity, which implies exponential convergence rates for $N \to \infty, \delta \to 0$.

6.3 Works which cite the article

The article "Mean-field Langevin dynamics and energy landscape of neural networks" is rather recent, since it was published in May 2019. It has been so far cited in the following four scientific studies:

Unbiased deep solvers for parametric PDEs - 24/07/2019
Development of several deep learning algorithms for approximating families of parametric PDE solutions. The proposed algorithms approximate solutions together with their gradients, which in the context of mathematical finance means that the derivative prices and hedging strategies are computed simultaneously. Having approximated the gradient of the solution one can combine it with a Monte-Carlo simulation to remove the bias in the deep network approximation of the PDE solution (derivative price). This is achieved by leveraging the Martingale Representation Theorem and combining the Monte Carlo simulation with the neural network. The resulting algorithm is robust with respect to quality of the neural network approximation and consequently can be used as a black-box in case only limited a priori information about the underlying problem is available. This is relevant as neural network based algorithms often require fair amount of tuning to produce satisfactory results.

The aim of the article is to develop algorithms that can be used as a black-box with only limited a priori information about the underlying problem. It focuses in particular on the problem of derivative pricing in high-dimensions with arbitrary payoff. It has been observed from the results in this article that neural networks provide efficient computational device for high dimensional problems. However, these algorithms are sensitive to the network architecture, parameters and distribution of training data. A fair amount of tuning is required to obtain good results.

The authors propose three classes of learning algorithms for simultaneously finding solutions and gradients to parametric families of PDEs:
- Projection solver (conditional expectation can be viewed as L2 projection operator)
- Martingale representation solver (MRS)
- Martingale control variates solver (MCV)

The application of deep neural networks trained with stochastic gradient descent algorithm to solve (or approximate solution) PDE is relatively new. PDEs provide an excellent test bed for neural networks approximation because there exists alternative solvers (Monte Carlo,...), there exist well developed theory for PDEs, and that knowledge can be used to tune algorithms. There has been also some important theoretical contributions. It has been proved that deep artificial neural networks approximate solutions to parabolic PDEs to an arbitrary accuracy without suffering from the curse of dimensionality. Furthermore, "Mean-field Langevin dynamics and energy landscape of neural networks" paper recently demonstrated that noisy gradient descent algorithm used for training of neural networks of certain form induces unique probability distribution function over the parameter space which minimizes learning. This means that there are theoretical guarantees for the approximation of (parabolic) PDEs with neural networks trained by noisy gradient methods alleviating the course of dimensionality. An important application of deep PDE solvers is that one can in fact approximate the parametric family of solutions of PDEs.

Mean-field Langevin System, Optimal Control and Deep Neural Networks - 03/10/2019
Optimal control deals with the problem of finding a control law for a given system such that a certain optimality criterion is achieved. A control problem includes a cost functional that is a function of state and control variables. An optimal control is a set of differential equations describing the paths of the control variables that minimize the cost function. In this paper, a regularized relaxed optimal control problem is studied and, in particular, the authors are concerned with the case where the control variable is of large dimension. They introduce a system of mean-field Langevin equations, the invariant measure of which is shown to be the optimal control of the initial problem under mild conditions. Therefore, this system of processes can be viewed as a continuous-time numerical algorithm for computing the optimal control. As an application, this result endorses the solvability of the stochastic gradient descent algorithm for a wide class of deep neural networks.

This paper revisits the classical optimal control problem, that is :
$\inf_\alpha V^0(\alpha)$, where $V^0(\alpha) := \int_0^T L\left(t, X_t^\alpha, \alpha_t\right) dt + G\left(X_T^\alpha\right)$ and $X_t = x_0 + \int_0^t \phi\left(r, X_r^\alpha, \alpha_r\right) dr$

In particular, it aims at providing a feasible algorithm for solving such problem when the dimensions of the state X and of the control alpha are both large. It has been more than half a century since the discovery of Pontryagin's maximum principle, which states that in order to be an optimal control to the problem, alpha* needs to satisfy the forward-backward ODE system:

$$\begin{cases} \alpha_t^* = \operatorname{argmin}_a H\left(t, X_t^*, a, P_t^*\right), \quad \text{where} \quad H(t,x,a,p) := L(t,x,a) + p \cdot \phi(t,x,a) \\ X_t^* = x_0 + \int_0^t \phi\left(r, X_t^*, \alpha_r^*\right) dr \\ P_t^* = \nabla_x G\left(X_T^*\right) + \int_0^t \nabla_x H\left(r, X_r^*, \alpha_r^*, P_r^*\right) dr \end{cases}$$

However, like other gradient-descent type algorithms, it would converge to a local minimizer, since Pontryagin's maximum principle is only a necessary 1rst-order condition. One may attempt to put a convexity condition on the coefficients in order to ensure the local minimizer to be the global one. However, this usually urges X to be linear in alpha (so the function phi needs to be linear in (x; a)), which largely limits the application of this method. In order to go beyond the convex case for the optimal control problem, it is natural to recall how the Langevin equation helps to approximate the solution of the non-convex optimization on the real space. Given a function F not necessarily convex, we know that under some mild conditions the unique invariant measure of the following Langevin equation

$\mathrm{d}\Theta_s = -\dot{F}\left(\Theta_s\right) ds + \sigma dW_s$ is the global minimizer of the regularized optimization:

$\min_{\nu \in \mathscr{P}} \int_{\mathbb{R}^m} F(a)\nu(da) + \frac{\sigma^2}{2} \operatorname{Ent}(\nu)$ where W is the Brownian motion, P is the space of probability measures and the regularizer Ent is the relative entropy with respect to the Lebesgue measure. Moreover, the marginal law of the process converges to its invariant measure. As analyzed in the recent paper "Mean-field Langevin dynamics and energy landscape of neural networks", this result is basically due to the fact that the function $\nu \mapsto \int F(a)\nu(da)$

is convex (indeed linear). In the present paper the authors wish to apply a similar regularization to the optimal control problem. In order to do that they recall the relaxed formulation of the control problem. Instead of controlling the process alpha, they will seek to control the flow of laws $(\nu_t)_{t\in[0,T]}$

Then the controlled process reads $\mathrm{X}_t = x_0+\int_0^t \int_{\mathbb{R}^m} \phi\left(r, X_r, a\right) \nu_r(da)dr$ and they aim at minimizing $\inf_\nu V(\nu), \quad \text{where} \quad V(\nu) := \int_0^T \int L\left(t, X_t, a\right) \nu_t(da)dt - G\left(X_T\right)$

Furthermore, they add the relative entropy as a regularizer, and focus on the regularized optimization: $\inf_\nu V^\sigma(\nu), \quad \text{where} \quad V^\sigma(\nu) := V(\nu) + \frac{\sigma^2}{2} \int_0^T \operatorname{Ent}\left(\nu_t\right) dt$

Mean Field Analysis of Neural Networks: A Law of Large Numbers – (First version in 05/2018, and last version 11/11/2019)

This paper analyzes one-layer neural networks in the asymptotic regime of simultaneously large network sizes and large numbers of stochastic gradient descent training iterations. The authors prove that the empirical distribution of the neural network parameters converges to the solution of a nonlinear partial differential equation. The proof relies upon weak convergence analysis for interacting particle systems. The result can be considered a law of large numbers for neural networks when both the network size and the number of stochastic gradient descent steps grow to infinity.

Mean Field Analysis of Neural Networks: A Law of Large Numbers – (First version in 05/2018, and last version 11/11/2019)

Analysis of deep neural networks and neural ODE models that are trained with stochastic gradient algorithms (connections between deep learning and controlled ODEs). The authors identify the connections between high-dimensional data-driven control problems, deep learning (deep neural network models) and theory of statistical sampling. They demonstrate how all of them are fundamentally intertwined. Optimal (relaxed) control perspective on deep learning tasks provides new insights, with a solid theoretical foundation. In particular, the powerful idea of relaxed control on generalized solutions of problems of calculus of variations, paves the way for efficient algorithms used in the theory of statistical sampling. Indeed, the task of learning the optimal weights in deep neural networks is viewed as a sampling problem, as shown in “Mean-field Langevin dynamics and energy landscape of neural networks” paper in the case of one-hidden layer network and in the case of ensembles of deep neural networks with the stochastic gradient algorithm.

In particular, the authors derive and study a mean-field (over-damped) Langevin algorithm for solving relaxed data-driven control problems. They prove convergence of flow of measures induced by the mean-field Langevin dynamics to invariant measure that minimized relaxed control problem. A key step in the analysis is to derive Pontryagin’s optimality principle for data-driven relaxed control problems. Subsequently, they study uniform-in-time propagation of chaos of time-discretized Mean-Field (overdamped) Langevin dynamics, and derive explicit convergence rate in terms of the learning rate, the number of particles/model parameters and the number of iterations of the gradient algorithm. In addition, they study the error arising when using a finite training data set and thus provide quantitative bounds on the generalization error. Crucially, the obtained rates are dimension-independent. This is possible by exploiting the regularity of the model with respect to the measure over the parameter space (relaxed control).

To work out some of their main results, the authors need the existence of the invariant measures and the required integrability and regularity for solutions to the Kolmogorov–Fokker–Planck equations proved in “Mean-field Langevin dynamics and energy landscape of neural networks” paper.

7 Python implemented animations to visualize the link between Langevin and Fokker-Planck dynamics, and the convergence of the flow of measures

Please refer to the animations and the Python code that have been shared through Google Drive :

```
https://drive.google.com/open?id=1gCvBL2QoHsF6zCKWINSiqGRcDl0JVLBX
```

In order to illustrate the correspondence between the Langevin equation describing the movement, the behavior of a single particle, and the Fokker-Planck equation describing the dynamics of a flow of measure, we decided to write a script in Python that generates animations of those dynamics given an energy function. The movement the particle following Langevin equation is not deterministic, thus we need to use the language of probability in order to describe the distribution of the particle at each time t (density of the random variable that describe the position of the particle). The Fokker-Planck equation describes just that : the evolution of this random variable. Those are two equations that describe the same thing, the same physical process from different points of view.

Those animations show indeed a convergence of the flow of measure (density) towards an invariant measure.