Mean-field Langevin dynamics in the energy landscape of neural networks[1]

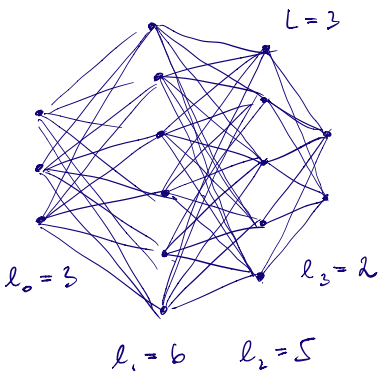
David Šiška[2]

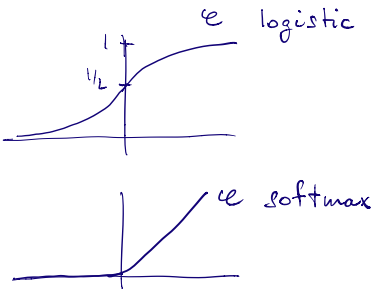
Joint work with Kaitong Hu[3], Zhenjie Ren[4] and Lukasz Szpruch[1]

Stochastic Analysis Seminar, University of Oxford
3rd June 2019

[1]https://arxiv.org/abs/1905.07769
[2]Universitiy of Edinburgh
[3]CMAP École Polytechnique
[4]CEREMADE, Université Paris-Dauphine

Neural networks





We are told these, but **much** bigger, will run everything...

Neural networks

. . . because they work really well for:

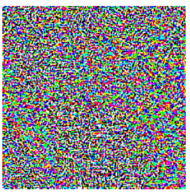
i) image recognition, see e.g. Huang et. al. [9],
ii) speech recognition, e.g. Dahl et. al. [3],
iii) numerical solution to PDEs, e.g. Vidales et. al. [15],
iv) dynamic hedging in finance, e.g. [1],
v) . . .

Until they don't

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

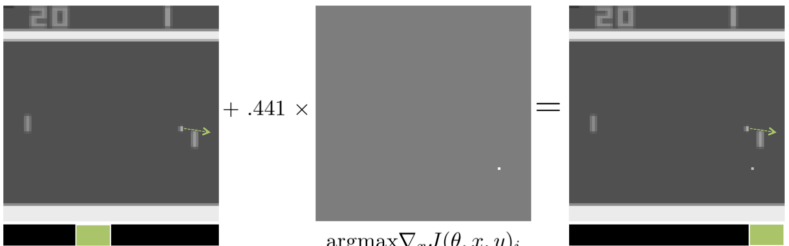
$=$

$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

From Goodfellow et. al. [4].

Until they don't



From Huang et. al. [10].

What is a neural net?

Parametric description of a function.

Fix

i) an *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$,

ii) number of layers $L \in \mathbb{N}$,

iii) the size of input to each layer k given by $l_k \in \mathbb{N}$, $k = 0, \ldots, L-1$,

iv) the size of the output layer $l_L \in \mathbb{N}$,

v) the space of parameters

$$\Pi = (\mathbb{R}^{l^1 \times l^0} \times \mathbb{R}^{l^1}) \times (\mathbb{R}^{l^2 \times l^1} \times \mathbb{R}^{l^2}) \times \cdots \times (\mathbb{R}^{l^L \times l^{L-1}} \times \mathbb{R}^{l^L}),$$

vi) the network *parameters*

$$\Psi = ((\alpha^1, \beta^1), \ldots, (\alpha^L, \beta^L)) \in \Pi.$$

The neural network

$$\Psi = ((\alpha^1, \beta^1), \ldots, (\alpha^L, \beta^L)) \in \Pi$$

now defines a function $\mathcal{R}\Psi : \mathbb{R}^{l^0} \to \mathbb{R}^{l^L}$ given recursively, for $x_0 \in \mathbb{R}^{l^0}$, by $z_0 \in \mathbb{R}^{l^0}$, by

$$(\mathcal{R}\Psi)(z^0) = \alpha^L z^{L-1} + \beta^L\,, \quad z^k = \varphi^{l^k}(\alpha^k z^{k-1} + \beta^k)\,, k = 1, \ldots, L-1\,.$$

Here $\varphi^{l_k} : \mathbb{R}^{l_k} \to \mathbb{R}^{l_k}$ is given, for $z = (z_1, \ldots, z_{l_k})^\top \in \mathbb{R}^{l_k}$, by $\varphi^{l_k}(z) = (\varphi(z_1), \ldots, \varphi(z_l))^\top$.

Example: One-hidden-layer network

Let $l_0 = d$, let $l_1 = n$, let $\beta^2 = 0 \in \mathbb{R}$, $\beta^1 = 0 \in \mathbb{R}^n$, $\alpha^1 \in \mathbb{R}^{n \times d}$. We will denote, for $i \in \{1, \dots, l^0\}$, its i-th row by $\alpha_i^1 \in \mathbb{R}^{1 \times d}$. Let $\alpha^2 = (\frac{c_1}{n}, \cdots, \frac{c_n}{n})^\top$, where $c_i \in \mathbb{R}$. The neural network is $\Psi^n = \Big((\alpha^1, \beta^1), (\alpha^2, \beta^2)\Big)$.

For $z \in \mathbb{R}^{l^0}$, its reconstruction can be written as

$$(\mathcal{R}\Psi^n)(z) = \alpha^2 \varphi^{l^1}(\alpha^1 z) = \frac{1}{n} \sum_{i=1}^{n} c_i \varphi(\alpha_i^1 \cdot z)\,.$$

Universal approximation theorem

If an activation function φ is bounded, continuous and non-constant, then for any compact set $K \subset \mathbb{R}^d$ the set

$$\Big\{ (\mathcal{R}\Psi) : \mathbb{R}^d \to \mathbb{R} : (\mathcal{R}\Psi) \text{ given above} \\ \text{with } L = 2 \text{ for some } n \in \mathbb{N}, \alpha_j^2, \beta_j^1 \in \mathbb{R}, \alpha_j^1 \in \mathbb{R}^d, j = 1, \ldots, n \Big\}$$

is dense in the space of continuous functions from K to $\mathbb{R}$. See e.g. Hornik [8, Theorem 2].

PDE approximation without the curse of dimensionality I

Consider

$$\begin{cases} \partial_t v + \mathrm{tr}[a\,\partial_x^2 v] + b\partial_x v = 0 \ \text{ in } [0,T)\times\mathbb{R}^d\,, \\ v(T,\cdot) = g \ \text{ on } \mathbb{R}^d\,, \end{cases}$$

where $a(x) = \frac{1}{2}\mathrm{diag}(x)\sigma\,[\mathrm{diag}(x)\sigma]^\top$ and $b(x) = \mathrm{diag}(x)\mu$. Let $(B_t)_{t\in[0,T]}$ be an $\mathbb{R}^{d'}$-valued Wiener process. The SDE arising in the Feynman–Kac representation for $v(t,x)$ is

$$dX_t^i = X_t^i\mu^i\,dt + X_t^i\sum_{j=1}^{d'}\sigma^{ij}\,dB_t^j\,,\ \ t\in[t,T]\,, X_t = x$$

and its solution is

$$X_T^i = x^i\exp\left[\left(\mu^i - \frac{1}{2}\sum_{j=1}^{d'}(\sigma^{ij})^2\right)(T-t) + \sum_{j=1}^{d'}\sigma^{ij}(B_T^j - B_t^j)\right] := \mathcal{W}_t^i x^i\,.$$

PDE approximation without the curse of dimensionality II

One-hidden-layer NN denoted Φ s.t. $g(x) = (\mathcal{R}\Phi)(x)$. Say

$$\Phi = \Big((\mathcal{W}_1^\Phi, \mathcal{B}_1^\Phi), (\mathcal{W}_2^\Phi, \mathcal{B}_2^\Phi)\Big) \in \big(\mathbb{R}^{1\times d} \times \mathbb{R}^1\big) \times \big(\mathbb{R}^{1\times 1} \times \mathbb{R}^1\big)$$

so that $(\mathcal{R}\Phi)(x) = \mathcal{W}_2^\Phi \mathbf{a}(\mathcal{W}_1^\Phi x + \mathcal{B}_1^\Phi) + \mathcal{B}_2^\Phi$. Further let us define

$$\mathcal{W}_1^\Psi := \underbrace{\operatorname{diag}(\mathcal{W}_1^\Phi, \dots, \mathcal{W}_1^\Phi)}_{\in \mathbb{R}^{N\times Nd}} \underbrace{\begin{pmatrix} \mathcal{W}_1 \\ \vdots \\ \mathcal{W}_N \end{pmatrix}}_{\in \mathbb{R}^{Nd\times d}} \in \mathbb{R}^{N\times d}, \quad \mathcal{B}_1^\Psi := \begin{pmatrix} \mathcal{B}_1^\Phi \\ \vdots \\ \mathcal{B}_1^\Phi \end{pmatrix} \in \mathbb{R}^N,$$

$$\mathcal{W}_2^\Psi := \frac{1}{N}(\mathcal{W}_2^\Phi, \dots, \mathcal{W}_2^\Phi) \in \mathbb{R}^{1\times N}, \quad \mathcal{B}_2^\Psi := B_2^\Phi \in \mathbb{R}^1$$

and

$$\Psi = \Big((\mathcal{W}_1^\Psi, \mathcal{B}_1^\Psi), (\mathcal{W}_2^\Psi, \mathcal{B}_2^\Psi)\Big) \in \big(\mathbb{R}^{N\times d} \times \mathbb{R}^N\big) \times \big(\mathbb{R}^{1\times N} \times \mathbb{R}^1\big).$$

PDE approximation without the curse of dimensionality III

Then for any $x \in \mathbb{R}^d$ we have that

$$
\begin{aligned}
(\mathcal{R}\Psi)(x) &= \mathcal{W}_2^{\Psi}\mathbf{A}_N\big(\mathcal{W}_1^{\Psi}x + \mathcal{B}_1^{\Psi}\big) + \mathcal{B}_2^{\Psi} \\
&= \frac{1}{N}(\mathcal{W}_2^{\Phi},\dots,\mathcal{W}_2^{\Phi})\mathbf{A}_N \begin{pmatrix} \mathcal{W}_1^{\Phi}\mathcal{W}_1 x + \mathcal{B}_1^{\Phi} \\ \vdots \\ \mathcal{W}_1^{\Phi}\mathcal{W}_N x + \mathcal{B}_1^{\Phi} \end{pmatrix} + \mathcal{B}_2^{\Phi} \\
&= \frac{1}{N}\sum_{k=1}^{N}\Big(\mathcal{W}_2^{\Phi}\mathbf{a}\big(\mathcal{W}_1^{\Phi}\mathcal{W}_k x + \mathcal{B}_1^{\Phi}\big) + \mathcal{B}_2^{\Phi}\Big) \\
&= \frac{1}{N}\sum_{k=1}^{N}\Big(\mathcal{W}_2^{\Phi}\mathbf{a}\big(\mathcal{W}_1^{\Phi}X_{T,k} + \mathcal{B}_1^{\Phi}\big) + \mathcal{B}_2^{\Phi}\Big) = \frac{1}{N}\sum_{k=1}^{N}\big(\mathcal{R}\Phi\big)(X_{T,k}) \\
&= \frac{1}{N}\sum_{k=1}^{N} g(X_{T,k}) \approx \mathbb{E}[g(\mathcal{W}_t x)] = v(t,x)\,.
\end{aligned}
$$

See series of works by Grohs, Hornung, Jentzen and von Wurstemberger [5] and Jentzen, Salimova and Welti [11].

PDE approximation without the curse of dimensionality IV

Note for later that

$$(\mathcal{R}\Psi)(x) = \frac{1}{N}\sum_{k=1}^{N} (\mathcal{R}\Phi)(\mathcal{W}_k x)$$
$$= \int_{\mathbb{R}^d} (\mathcal{R}\Phi)(y\, x) m^N(dy)\,,$$

where

$$m^N := \frac{1}{N}\sum_{k=1}^{N} \delta_{\mathcal{W}_k}\,.$$

In fact

$$v(t,x) = \int_{\mathbb{R}^d} (\mathcal{R}\Phi)(y\, x) m^*(dy) \quad \text{where } m^* \text{ is the law of } X_T^{t,x}.$$

What is understood in deep learning

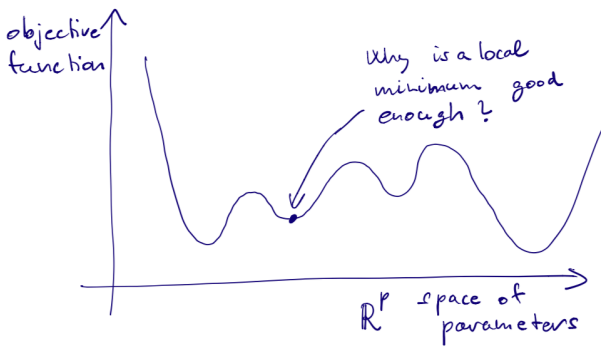
i) Representation theorems for various settings,

ii) Deep networks are a way to reduce number of parameters ,

iii) . . .

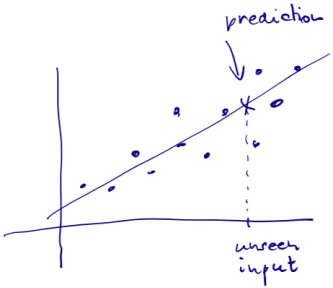
What is not so well understood in deep learning

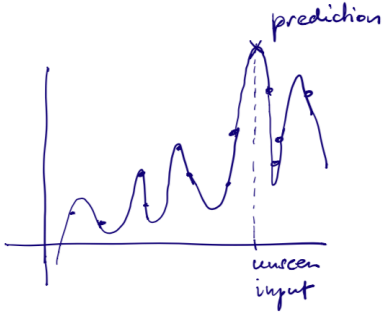
i) Why does supervised learning (nonconvex optimization) work?



What is not so well understood in deep learning

ii) How come massively over-parametrized models generalize well?





See Hastie, Montanari, Rosset and Tibshirani [7].

Supervised learning

So we have mathematical theory that shows that *optimal* parameters for the network exists. How to find them?

Supervised learning:

i) $\Phi : \mathbb{R} \to \mathbb{R}^+$ given, convex, e.g. $\Phi(x) = |x|^2$

ii) sample learning data from measure $\nu \in \mathcal{P}(\mathbb{R} \times \mathbb{R}^D)$ i.e. "big data"

iii) aim is to find optimal network parameters w.r.t. Φ.

Non-covex minimization problem

With $\hat{\varphi}(x,z) = \beta\varphi(\alpha \cdot z)$ for $x = (\alpha, \beta) \in (\mathbb{R} \times \mathbb{R}^D)^n$, we should minimize,

$$(\mathbb{R} \times \mathbb{R}^D)^n \ni x \mapsto \underbrace{\int_{\mathbb{R}\times\mathbb{R}^D} \Phi\left(y - \frac{1}{n}\sum_{i=1}^{n} \hat{\varphi}(x^i, z)\right) \nu(dy, dz)}_{=:F(x)} + \frac{\bar{\sigma}^2}{2} \underbrace{|x|^2}_{=:U(x)},$$

which is non-convex.

Gradient descent with "learning rate" $\tau > 0$:

$$x_{k+1}^i = x_k^i - \tau \nabla_{x^i} \left[F(x_k) + \frac{\bar{\sigma}^2}{2} |x_k^i|^2 \right], \quad i = 1, \ldots, n.$$

Here $x^i = (\alpha^i, \beta^i) \in \mathbb{R} \times \mathbb{R}^D$.

Approximation with gradient descent

In practice noisy, regularized, gradient descent algorithms are used:

$$x^i_{k+1} = x^i_k + \tau \int_{\mathbb{R}\times\mathbb{R}^D} \dot{\Phi}\left(y - \frac{1}{n}\sum_{j=1}^n \hat{\varphi}(x^j_k, z)\right) \nabla_{x^i}\hat{\varphi}(x^i_k, z)\, \nu(dy, dz) - \frac{\bar{\sigma}^2}{2}\nabla_{x^i} U(x^i_k) + \bar{\sigma}\sqrt{\tau}\xi^i_k\,,$$

where $(y_k, z_k)_{k\in\mathbb{N}}$ are i.i.d. samples from ν and ξ^i_k are i.i.d. samples from $N(0, I_d)$.

Donsker's invariance theorem tells us that with $W^{n,i}_t := n^{-1/2}\sum_{k=1}^{\lfloor nt\rfloor}\xi^i_k$ we have $W^{(n,i)}_t \implies W_t$ as $n\to\infty$ and the limiting dynamics is (after re-scaling)

$$dX^i_t = \left[\int_{\mathbb{R}\times\mathbb{R}^D} \dot{\Phi}\left(y - \frac{1}{n}\sum_{j=1}^n \hat{\varphi}(X^j_t, z)\right)\nabla_{x^i}\hat{\varphi}(X^i_t, z)\,\nu(dy, dz) - \frac{\bar{\sigma}^2}{2}\nabla_{x^i}U(X^i_t)\right]dt + \sigma dW^i_t\,,$$

Mean-field limit and convexity

Assume that x^i are i.i.d. samples from some measure $m \in \mathcal{P}(\mathbb{R}^d)$.

Due to law of large numbers, for each fixed $z \in \mathbb{R}^D$ we have

$$\frac{1}{n}\sum_{i=1}^{n} \hat{\varphi}(x^i, z) \to \int_{\mathbb{R}^d} \hat{\varphi}(x, z)\, m(dx) \;\text{ as }\; n \to \infty\,.$$

The search for the optimal measure $m^* \in \mathcal{P}(\mathbb{R}^d)$ amounts to minimizing

$$\mathcal{P}(\mathbb{R}^d) \ni m \mapsto \int_{\mathbb{R}\times\mathbb{R}^D} \Phi\left(y - \int_{\mathbb{R}^d} \hat{\varphi}(x, z)\, m(dx)\right)\nu(dy, dz) =: F(m),$$

which is convex as long as Φ is.

Observed in the pioneering works of Mei, Misiakiewicz and Montanari [12], Chizat and Bach [2] as well as Rotskoff and Vanden-Eijnden [14].

Propagation of chaos

On the level of the particle system

$$dX_t^i = \Bigg[\int_{\mathbb{R}\times\mathbb{R}^D} \dot{\Phi}\Big(y - \frac{1}{n}\sum_{j=1}^{n} \hat{\varphi}(X_t^j, z) \Big) \nabla\hat{\varphi}(X_t^i, z)\, \nu(dy, dz) - \frac{\bar{\sigma}^2}{2} \nabla U(X_t^i) \Bigg] dt + \sigma dW_t^i \,,$$

we expect to have, as $n \to \infty$,

$$\begin{cases} dX_t = -\left(D_m F(m_t, X_t) + \dfrac{\sigma^2}{2} \nabla U(X_t) \right) dt + \sigma dW_t \quad t \in [0, \infty) \\ m_t = \mathrm{Law}(X_t) \quad t \in [0, \infty) \,. \end{cases}$$

Fokker–Planck

$$\partial_t m = \nabla \cdot \left(\Big(D_m F(m, \cdot) + \frac{\sigma^2}{2} \nabla U \Big) m + \frac{\sigma^2}{2} \nabla m \right) \text{ on } (0, \infty) \times \mathbb{R}^d \,.$$

Measure derivatives

Example: If $x, y \in \mathbb{R}^d$ then $\nabla_x \langle x, y \rangle = y$.

Example: $v(m) = \int_{\mathbb{R}^d} f(x)\, m(dx) = \langle m, f \rangle$. So perhaps we want $\frac{\delta v}{\delta m} = f$?

Definition 1 (**Functional derivative**)

For $V : \mathcal{P} \to \mathbb{R}$ we say the *functional derivative* exists if there is a continuous map $\frac{\delta V}{\delta m} : \mathcal{P} \times \mathbb{R}^d \to \mathbb{R}$ such that for any $m, m' \in \mathcal{P}$

$$\lim_{s \searrow 0} \frac{V((1-s)m + sm') - V(m)}{s} = \int_{\mathbb{R}^d} \frac{\delta V}{\delta m}(m, y) d(m' - m)(y)\,.$$

Indeed for $v(m) = \langle m, f \rangle$ we have

$$\lim_{s \searrow 0} \frac{\langle (1-s)m + sm \rangle - \langle m, f \rangle}{s} = \langle m' - m, f \rangle = \int_{\mathbb{R}^d} f(y)\, d(m' - m)(y)\,.$$

So $\frac{\delta v}{\delta m} = f$ (up to a constant, normalize so that functional derivative integrates to 0).

Measure derivatives

Definition 2 (**Intrinsic derivative**)

For $V : \mathcal{P}_2 \to \mathbb{R}$ we say the *intrinsic derivative* exists if $\frac{\delta V}{\delta \mu} : \mathcal{P}_2 \times \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable in the 2nd variable and we say the function $D_m V : \mathcal{P}_2 \times \mathbb{R}^d \to \mathbb{R}$ given by

$$D_m V(m, x) := \nabla_x \frac{\delta V}{\delta m}(m, x)$$

is the intrinsic derivative.

Indeed for $v(m) = \langle m, f \rangle$ we have

$$D_m v(m, x) = \nabla_x f(x)\,.$$

Energy functional

Fix a Gibbs measure g:

$$g(x) = e^{-U(x)} \text{ with } U \text{ s.t. } \int_{\mathbb{R}^d} e^{-U(x)}\, dx = 1\,.$$

Define the relative entropy H for $m \in \mathcal{P}(\mathbb{R}^d)$ as:

$$H(m) := \begin{cases} \displaystyle\int_{\mathbb{R}^d} m(x) \log\left(\frac{m(x)}{g(x)}\right) dx \text{ if } m \text{ is a.c. w.r.t. Lebesgue measure}\,, \\ \infty \text{ otherwise}\,. \end{cases}$$

We will study $V^\sigma(m) := F(m) + \frac{\sigma^2}{2} H(m)$.

We have $\frac{\delta H}{\delta m}(m) = \log(m) - U$ and so

$$m \nabla \frac{\delta H}{\delta m}(m) = \nabla m - m \nabla U$$

which is the term in the Fokker–Planck due to the noise.

Assumptions I

Assumption 1

$F \in \mathcal{C}^1$ is convex and bounded from below.

Assumption 2

The function $U : \mathbb{R}^d \to \mathbb{R}$ belongs to C^∞. Further,

i) *there exist constants $C_U > 0$ and $C'_U \in \mathbb{R}$ such that*

$$\nabla U(x) \cdot x \geq C_U |x|^2 + C'_U \quad \text{for all} \quad x \in \mathbb{R}^d .$$

ii) *∇U is Lipschitz continuous.*

Convergence when $\sigma \searrow 0$

Proposition 3

Assume that F is continuous in the topology of weak convergence. Then the sequence of functions $V^\sigma = F + \frac{\sigma^2}{2} H$ converges in the sense of Γ-convergence to F as $\sigma \searrow 0$. In particular, given a minimizer $m^{*,\sigma}$ of V^σ, we have

$$\limsup_{\sigma \to 0} F(m^{*,\sigma}) = \inf_{m \in \mathcal{P}_2(\mathbb{R}^d)} F(m).$$

Proof outline: To get $\liminf_{\sigma_n \to 0} V^{\sigma_n}(m_n) \geq F(m)$ use l.s.c. of entropy.

To get $\limsup_{\sigma_n \to 0} V^{\sigma_n}(m_n) \leq F(m)$ smooth with heat kernel and use assumption of quadratic growth of U. ■

Characterization of the minimizer

Proposition 4

Under Assumption 1 and 2, the function V^σ has a unique minimizer $m^* \in \mathcal{P}_2(\mathbb{R}^d)$ which is absolutely continuous with respect to Lebesgue measure and satisfies

$$\frac{\delta F}{\delta m}(m^*, \cdot) + \frac{\sigma^2}{2}\log(m^*) + \frac{\sigma^2}{2}U \quad \textit{is a constant, } m^* - a.s.$$

On the other hand if $m' \in \mathcal{I}_\sigma$ where

$$\mathcal{I}_\sigma := \left\{ m \in \mathcal{P}(\mathbb{R}^d) : \frac{\delta F}{\delta m}(m, \cdot) + \frac{\sigma^2}{2}\log(m) + \frac{\sigma^2}{2}U \quad \textit{is a constant} \right\}$$

then $m' = \arg\min_{m \in \mathcal{P}(\mathbb{R}^d)} V^\sigma$.

Proof outline: Step 1: Sublevel sets of the entropy are compact so consider, for some fixed $\bar{m}$ s.t. $V(\bar{m}) < \infty$,

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

Since V is l.s.c. it attains its minimum on $\mathcal{S}$, say m^* so $V(m^*) \leq V(m)$ for all $m \in \mathcal{S}$.

Note that $\bar{m} \in \mathcal{S}$. If $m \notin \mathcal{S}$ then

$$V(m^*) \leq V(\bar{m}) \leq \frac{\sigma^2}{2} H(m) + \inf_{m' \in \mathcal{P}(\mathbb{R}^d)} F(m') \leq V(m)$$

so m^* is global minimum of V. Since V is strictly convex it is unique.

Step 2: Assume $m^* \in \mathcal{I}_\sigma$ and show that for any $\varepsilon > 0$ and $m \in \mathcal{P}(\mathbb{R}^d)$ you have

$$\frac{V((1-\varepsilon m^*) + \varepsilon m) - V(m^*)}{\varepsilon}$$
$$\geq \int_{\mathbb{R}^d} \left(\frac{\delta F}{\delta m}(m^*, \cdot) + \frac{\sigma^2}{2} \log m^* + \frac{\sigma^2}{2} U \right) (m - m^*)(dx) = 0 \,.$$

Step 3: Assume m^* is the minimizer of V. Let $m \in \mathcal{P}(\mathbb{R}^d)$ be arbitrary.

Use definition of linear functional derivative to show that

$$0 \leq \int_{\mathbb{R}^d} (m - m^*)(x) \left(\frac{\delta F}{\delta m}(m^*, x) + \frac{\sigma^2}{2} \log(m^*(x)) + \frac{\sigma^2}{2} U(x) + \frac{\sigma^2}{2} \right) dx.$$

■

Connection to gradient flow

If $m^* \in \mathcal{I}_\sigma$ then

$$\frac{\delta F}{\delta m}(m^*, \cdot) + \frac{\sigma^2}{2} \log(m^*) + \frac{\sigma^2}{2} U \text{ is a constant, } m^* - a.s.$$

and so (formally, apply ∇, multiply by m^*, apply $\nabla \cdot$)

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

and so it is (formally) the stationary solution of

$$\partial_t m = \nabla \cdot \left(\left(D_m F(m, \cdot) + \frac{\sigma^2}{2} \nabla U \right) m + \frac{\sigma^2}{2} \nabla m \right) \text{ on } (0, \infty) \times \mathbb{R}^d .$$

Mean-field Langevin equation

We see that if

$$\begin{cases} dX_t = -\left(D_m F(m_t, X_t) + \dfrac{\sigma^2}{2} \nabla U(X_t) \right) dt + \sigma dW_t \quad t \in [0, \infty) \\ m_t = \text{Law}(X_t) \quad t \in [0, \infty) \end{cases} \tag{1}$$

has a solution then $(m_t)_{t \geq 0}$ solves the Fokker–Planck equation arising from the 1st order condition i.e.

$$\partial_t m = \nabla \cdot \left(\left(D_m F(m, \cdot) + \frac{\sigma^2}{2} \nabla U \right) m + \frac{\sigma^2}{2} \nabla m \right) \text{ on } (0, \infty) \times \mathbb{R}^d .$$

Assumptions II

Assumption 5

Assume that the intrinsic derivative $D_m F : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ of the function $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ exists and satisfies the following conditions:

i) *$D_m F$ is bounded and Lipschitz continuous, i.e. there exists $C_F > 0$ such that for all $x, x \in \mathbb{R}^d$ and $m, m' \in \mathcal{P}_2(\mathbb{R}^d)$*

$$|D_m F(m, x) - D_m F(m', x')| \leq C_F \left(|x - x'| + \mathcal{W}_2(m, m') \right).$$

ii) *$D_m F(m, \cdot) \in \mathcal{C}^\infty(\mathbb{R}^d)$ for all $m \in \mathcal{P}(\mathbb{R}^d)$.*

iii) *$\nabla D_m F : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d \times \mathbb{R}^d$ is jointly continuous.*

Proposition 6

If Assumptions 2 and 5 hold and if $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *then:*

i) *the mean field Langevin SDE* (1) *has a unique strong solution,*

ii) *given* $m_0, m_0' \in \mathcal{P}_2(\mathbb{R}^d)$ *and denoting by* $(m_t)_{t\geq 0}, (m_t')_{t\geq 0}$ *the marginal laws of the corresponding solutions to* (1)*, we have for all* $t > 0$ *that there is a constant* $C > 0$ *such that*

$$\mathcal{W}_2(m_t, m_t') \;\leq\; C\mathcal{W}_2(m_0, m_0')\,.$$

Theorem 3

Let $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumption 2 and 5, we have for any $t > s > 0$

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

Proof outline: Follows from a priori estimates and regularity results on the nonlinear Fokker–Planck equation and the chain rule for flows of measures.

Convergence

Theorem 4

Let Assumption 1, 2 and 5 hold true and $m_0 \in \cup_{p>2}\mathcal{P}_p(\mathbb{R}^d)$. *Denote by* $(m_t)_{t\geq 0}$ *the flow of marginal laws of the solution to* (1). *Then, there exists an invariant measure of* (1) *equal to* $m^* := \operatorname{argmin}_m V^\sigma(m)$ *and*

$$\mathcal{W}_2(m_t, m^*) \to 0 \ \text{ as } \ t \to \infty\,.$$

Proof key ingredients: Tightness of $(m_t)_{t\geq 0}$, Lasalle's invariance principle, Theorem 3, HWI inequality.

Convergence, step 1: invariance

Let $S(t)[m_0] := m_t$, marginals of solution to (1) started from m_0.

From $m_0 \in \bigcup_{p>2} \mathcal{P}_p(\mathbb{R}^d)$ let

$$\omega(m_0) := \left\{ \mu \in \mathcal{P}_2(\mathbb{R}^d) : \exists (t_n)_{n\in\mathbb{N}} \text{ s.t. } \mathcal{W}_2(m_{t_n}, \mu) \to 0 \text{ as } n \to \infty \right\}.$$

Then

i) $\omega(m_0)$ is nonempty and compact,
ii) if $\mu \in \omega(m_0)$ then $S(t)[\mu] \in \omega(m_0)$ for all $t \geq 0$,
iii) if $\mu \in \omega(m_0)$ then for any $t \geq 0$ there exists μ' s.t. $S(t)[\mu'] = \mu$.

Convergence, step 1: invariance

Then: from i) $\implies$ there is $\tilde{m} \in \operatorname{argmin}_{m \in \omega(m_0)} V(m)$.

from iii) $\forall t > 0$ there is μ s.t. $S(t)[\mu] = \tilde{m}$ and by Theorem 3 for any $s > 0$ we get

$$V(S(t+s)[\mu]) \leq V(S(t)[\mu]) = V(\tilde{m})\,.$$

from ii) $S(t+s)[\mu] \in \omega(m_0)$ so $V(S(t+s)[\mu]) \geq V(\tilde{m})$. By Theorem 3

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

Due to the first order condition (Proposition 4) get $\tilde{m} = m^*$.

Convergence, step 2: HWI inequality

We want to show that if $m_{t_n} \to m^*$ then $V(m_{t_n}, m^*) \to 0$.

But $V = F + \frac{\sigma^2}{2} H$ and H only l.s.c. So we need to show that

$$\int_{\mathbb{R}^d} m^* \log(m^*)\, dx \geq \limsup_{n\to\infty} \int_{\mathbb{R}^d} m_{t_n} \log(m_{t_n})\, dx\,.$$

Convergence, step 2: HWI inequality

Otto, Villani [13, Theorem 3]:

Assume that $\nu(dx) = e^{-\Psi(x)}(dx)$ is a $\mathcal{P}_2(\mathbb{R}^d)$ measure s.t. $\Psi \in C^2(\mathbb{R}^d)$, there is $K \in \mathbb{R}$ s.t. $\partial_{xx}\Psi \geq K I_d$. Then for any $\mu \in \mathcal{P}(\mathbb{R}^d)$ absolutely continuous w.r.t. ν we have

$$H(\mu|\nu) \leq \mathcal{W}_2(\mu,\nu)\left(\sqrt{I(\mu|\nu)} - \frac{K}{2}\mathcal{W}_2(\mu,\nu)\right),$$

where I is the Fisher information:

$$I(\mu|\nu) := \int_{\mathbb{R}^d} \left|\nabla \log \frac{d\mu}{d\nu}(x)\right|^2 \mu(dx).$$

Convergence, step 2: HWI inequality

We thus have

$$\int_{\mathbb{R}^d} m_{t_n}\Big(\log(m_{t_n}) - \log(m^*)\Big)\,dx \leq \mathcal{W}_2(m_{t_n}, m^*)\Big(\sqrt{I_n} + C\mathcal{W}_2(m_{t_n}, m^*)\Big),$$

with

$$I_n := \mathbb{E}\left[\left|\nabla \log\left(m_{t_n}(X_{t_n})\right) - \nabla \log\left(m^*(X_{t_n})\right)\right|^2\right].$$

Need to show $\sup_n I_n < \infty$ (estimate on Malliavin derivative of the change of measure exponential).

Convergence, step 3

Have $m_{t_n} \to m^*$ for some $t_n \to \infty$. Moreover $t \mapsto V(m_t)$ is non-increasing so there is $c := \lim_{n\to\infty} V(t_n)$.

Use uniqueness of m^* and step 2 to show that any other sequence $V(m_{t_{n'}})$ converges to the same c, $\omega(m_0) = \{m^*\}$, so $\mathcal{W}_2(m_{t_{n'}}, m^*) \to 0$. ■

Assumption 7 (For exponential convergence)

Let $\sigma > 0$ be fixed and the mean-field Langevin dynamics (1) start from $m_0 \in \mathcal{P}_p(\mathbb{R}^d)$ for some $p > 2$. Assume that there are constants $C > 0$, $C_F > 0$ and $C_U > 0$ such that for all $x, x' \in \mathbb{R}^d$ and $m, m' \in \mathcal{P}_1(\mathbb{R}^d)$ we have

$$\begin{aligned} |D_m F(m,x) - D_m F(m',x')| &\le C_F\Big(|x-x'| + \mathcal{W}_1(m,m')\Big)\,, \\ |D_m F(m,0)| &\le C_F\Big(1 + \int_{\mathbb{R}^d} |y|\, m(\mathrm{d}y)\Big), \end{aligned} \tag{2}$$

$$\begin{aligned} (\nabla U(x) - \nabla U(x'))\cdot(x-x') &\ge C_U|x-x'|^2\,, \\ |\nabla U(x)| &\le C_U(1+|x|)\,, \end{aligned} \tag{3}$$

where the constants satisfy

$$\frac{\sigma^2}{2}(p-1) + 3C_F + \frac{\sigma^2}{2}|\nabla U(0)| - C_U\frac{\sigma^2}{2} < 0\,. \tag{4}$$

Exponential convergence

Theorem 5

Let Assumptions 1 and 7 hold true. Then

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

where $(m_t)_{t\geq 0}$ *is the flow of marginal laws of solution to* (1).

Proof outline: Use "integrated Lyapunov condition" from Hammersley, S. and Szpruch [6].

Main thing to show: for any $m \in \mathcal{P}(\mathbb{R}^d)$, that

$$\int_{\mathbb{R}^d} L(m,x)v(x)\,m(dx) \leq \frac{\sigma^2}{2}p(p-1) + pC_F + p\frac{\sigma^2}{2}|\nabla U(0)|$$
$$+ p\int_{\mathbb{R}^d}\Big[\frac{\sigma^2}{2}(p-1) + 3C_F + \frac{\sigma^2}{2}|\nabla U(0)| - C_U\frac{\sigma^2}{2}\Big]|x|^p\,m(dx)\,.$$

Particle approximation of m^*

Theorem 6

We assume that the 2nd order linear functional derivative of F exists, is jointly continuous in both variables and that there is $L > 0$ such that for any random variables η_1, η_2 such that $\mathbb{E}[|\eta_i|^2] < \infty$, $i = 1, 2$, it holds that

$$\mathbb{E}\left[\sup_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left|\frac{\delta F}{\delta m}(\nu, \eta_1)\right|\right] + \mathbb{E}\left[\sup_{\nu \in \mathcal{P}_2(\mathbb{R}^d)} \left|\frac{\delta^2 F}{\delta m^2}(\nu, \eta_1, \eta_2)\right|\right] \leq L \qquad (5)$$

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

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

Proof outline: Coupling argument.

Stochastic PDE

Go back to SGD:

$$x_{k+1}^i = x_k^i + \tau \dot{\Phi}\left(y_k - \frac{1}{n}\sum_{j=1}^n \hat{\varphi}(x_k^j, z_k)\right)\nabla\hat{\varphi}(x_k^i, z^k) - \frac{\bar{\sigma}^2}{2}\nabla U(x_k^i) + \bar{\sigma}\sqrt{\tau}\xi_k^i + \bar{\sigma}_0\sqrt{\tau}\chi_k\,,$$

where $(y_k, z_k)_{k\in\mathbb{N}}$ are i.i.d. samples from ν and χ_k, ξ_k^i are i.i.d. samples from $N(0, I_d)$.

Now χ represents *common noise* in the algorithm.

We would then need to consider stochastic PDE

$$dm_t = \nabla\cdot\left(\left(D_m F(m_t, \cdot) + \frac{\sigma^2}{2}\nabla U\right)m_t + \frac{\sigma^2 + \sigma_0^2}{2}\nabla m_t\right) + \sigma_0\, dB_t$$

on $(0,\infty)\times\mathbb{R}^d$.

Outlook

We have (nearly) full analysis of convergence of gradient descent algorithm for (some) deep networks.

i) Uniform-in-time propagation of chaos,
ii) Multiplicative noise in the dynamics,
iii) Other deep network architectures,
iv) Common noise case i.e. SPDE,
v) Design better algorithms based on understood theory: faster convergence, stability w.r.t. $\mathcal{W}_2$ metric etc.

References I

[1] Buehler, H., Gonon, L., Teichmann, J., and Wood, B. Deep hedging. *arXiv:1802.03042* (2018).

[2] Chizat, L., and Bach, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems* (2018), pp. 3040–3050.

[3] Dahl, G. E., Yu, D., Deng, L., and Acero, A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing 20*, 1 (2012), 30–42.

[4] Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv:1412.6572* (2014).

[5] Grohs, P., Hornung, F., Jentzen, A., and von Wurstemberger, P. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black–Scholes partial differential equations. *arXiv:1809.02362* (2018).

[6] Hammersley, W., Šiška, D., and Szpruch, L. McKean–Vlasov SDEs under measure dependent Lyapunov conditions. *arXiv:1802.03974* (2018).

[7] Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv:1903.08560* (2019).

[8] Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks 4*, 2 (1991), 251–257.

References II

[9] Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. *CVPR 1* (2017).

[10] Huang, S., Papernot, N., Goodfellow, I., Duan, Y., and Abbeel, P. Adversarial attacks on neural network policies. *arXiv:1702.02284* (2017).

[11] Jentzen, A., Salimova, D., and Welti, T. A proof that deep artificial neural networks overcome the curse of dimensionality in the numerical approximation of Kolmogorov partial differential equations with constant diffusion and nonlinear drift coefficients. *arXiv:1809.07321* (2018).

[12] Mei, S., Montanari, A., and Nguyen, P.-M. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences 115*, 33 (2018), E7665–E7671.

[13] Otto, F., and Villani, C. Generalization of an inequality by talagrand and links with the logarithmic sobolev inequality. *Journal of Functional Analysis 173* (2000), 361–400.

[14] Rotskoff, G. M., and Vanden-Eijnden, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error. *arXiv:1805.00915* (2018).

[15] Vidales, M. S., Šiška, D., and Szpruch, L. Martingale functional control variates via deep learning. *arXiv:1810.05094* (2018).