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

Lukasz Szpruch[2]

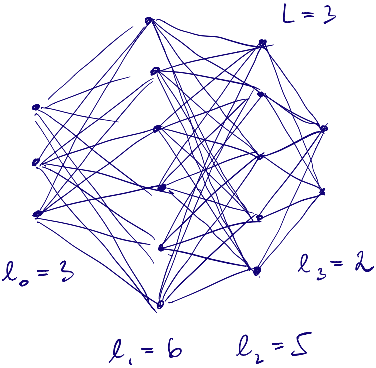
Joint work with Kaitong Hu[3], Zhenjie Ren[4] and David Šiška[2]

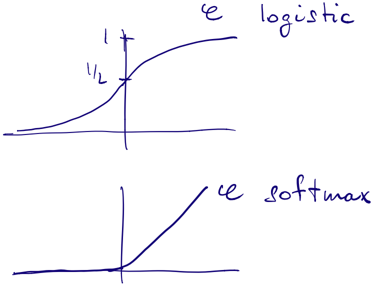
Stochastic Analysis Seminar, University of Oxford
3rd June 2019

[1]https://arxiv.org/abs/1905.07769

[2]Universitiy of Edinburgh
[3]CMAP École Polytechnique
[4]CEREMADE, Université Paris-Dauphine

Neural networks





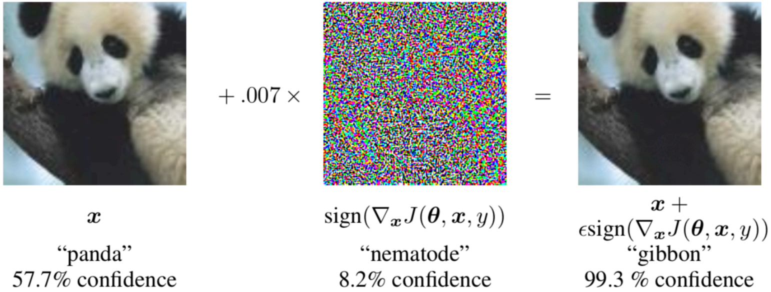
We are told these, but **much** bigger, will run everything. . .

Neural networks

. . . because they work really well for:

i) image recognition, see e.g. Huang et. al. [12],
ii) speech recognition, e.g. Dahl et. al. [5],
iii) numerical solution to PDEs, e.g. Vidales et. al. [19],
iv) dynamic hedging in finance, e.g. [1],
v) . . .

Until they don't



From Goodfellow et. al. [7].

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 $l_k \in \mathbb{N}$, $k = 0, \dots, L-1$,
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*

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

The neural network

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

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

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

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

Example: One-hidden-layer network

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

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

where 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 $\Psi^n = \Big((\alpha^1, \beta^1), (\alpha^2, \beta^2)\Big)$.

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 [11, 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

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

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. $g(x) = (\mathcal{R}\Phi)(x)$.

$$v(t,x) = E[g(\mathcal{W}_t x)] \approx \frac{1}{N}\sum_{k=1}^{N} g(\mathcal{W}_t^k x)\,.$$

See series of works by Grohs, Hornung, Jentzen and von Wurstemberger [8] and Jentzen, Salimova and Welti [13].

Note for later that

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

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

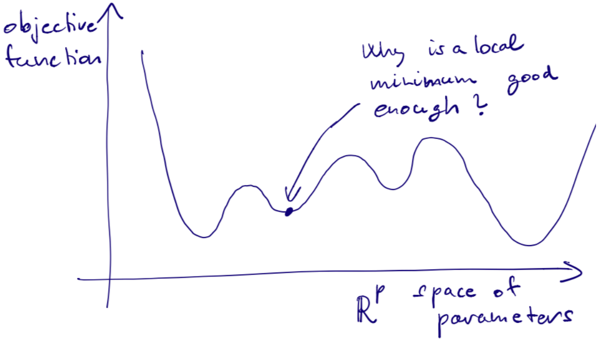
i) Representation theorems for various settings,

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

iii) . . .

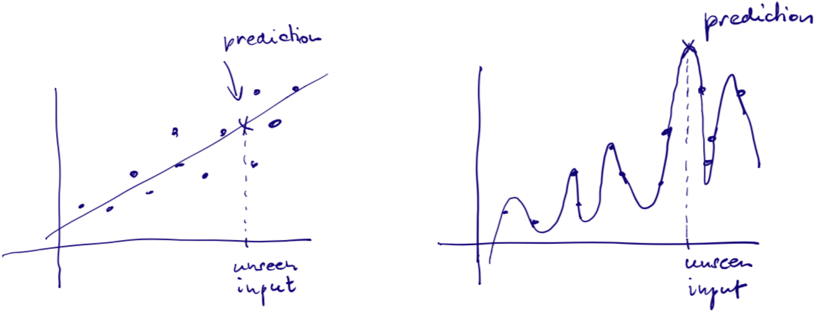
What is not so well understood in deep learning

i) Why gradient algorithms in non-convex optimization do the job?



What is not so well understood in deep learning

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



See Hastie, Montanari, Rosset and Tibshirani [10].

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" $\tau > 0$:

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

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:

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

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

Taking weak limit gives

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

Mean-field limit and convexity

Write

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

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) i.e

$$F((1-\alpha)m+\alpha m') \le (1-\alpha)F(m)+\alpha F(m') \;\text{ for all }\; \alpha\in[0,1]\,.$$

Observed in the pioneering works of Mei, Misiakiewicz and Montanari [14], Chizat and Bach [4] as well as Rotskoff and Vanden-Eijnden [17].

Derivation of MFLD I

$$F^N(x) = F\left(\frac{1}{N}\sum_{i=1}^{N}\delta_{x^i}\right) = \int_{\mathbb{R}^d}\Phi\Big(y - \frac{1}{N}\sum_{j=1}^{N}\hat{\varphi}(x^j, z)\Big)\,\nu(\mathrm{d}z, \mathrm{d}y).$$

Hence

$$\partial_{x^i}F^N(x^1, \ldots, x^N) = -\frac{1}{N}\int_{\mathbb{R}^d}\dot{\Phi}\Big(y - \frac{1}{N}\sum_{j=1}^{N}\hat{\varphi}(x^j, z)\Big)\nabla\hat{\varphi}(x^i, z)\nu(\mathrm{d}z, \mathrm{d}y)\,,$$

On the level of the particle system

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

Derivation of MFLD II

Then

$$\mathrm{d}X_t^i = -\Big(N\partial_{x_i}F^N(X_t^1,\dots,X_t^N) + \frac{\sigma^2}{2}\nabla U(X_t^i)\Big)\mathrm{d}t + \sigma\mathrm{d}W_t^i\,.$$

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

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

Fokker–Planck

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

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

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

Variational Perspective

Given a *potential* function $f : \mathbb{R}^d \to \mathbb{R}$ the overdamped Langevin dynamics (LD) reads

$$dX_t = -\nabla f(X_t)\mathrm{d}t + \sigma dW_t,$$

i) The solution to LD under mild conditions admits a unique invariant measure $m^{\sigma,*}$ with density

$$m^{\sigma,*}(x) = \frac{1}{Z}\exp\left(-\frac{2}{\sigma^2}f(x)\right), \forall x \in \mathbb{R}^d\,,\; Z := \int_{\mathbb{R}^d}\exp\left(-\frac{2}{\sigma^2}f(x)\right)\,\mathrm{d}x\,.$$

ii) The dynamic LD can be viewed as the path of a randomised continuous time gradient descent algorithm.

Note $m^{\sigma,*}$ is the unique minimiser of the free energy function

$$V^\sigma(m) := \int_{\mathbb{R}^d} f(x)m(dx) + \frac{\sigma^2}{2}H(m)$$

over all probability measure m,

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

$$dX_t = -\left(\nabla_x \frac{\delta F}{\delta m}(m_t, X_t) + \frac{\sigma^2}{2} \nabla U(X_t)\right) dt + \sigma dW_t \quad t \in [0, \infty)\,.$$

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 (existence of unique minimiser): 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 (sufficient condition): 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 (necessary condition): similar to step 2

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

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

and

$$m^*(x) = \frac{1}{Z} \exp\left(-\frac{2}{\sigma^2} \left(\frac{\delta F}{\delta m}(m^*, x) + U(x) \right) \right),$$

Mean-field Langevin equation

We see that if

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

has a solution then $(m_t)_{t\geq 0}$ solves the Fokker–Planck equation

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

Key challenges in studying invariant measure(s)

- Drift not of convolutional form Carillo, McCann Vilani [2] Otto [15], Tugaut [18]
- To establish the link with optimisation need result to hold for all σ Bogachev, Roeckner, Shaposhnikov [**?**] and Eberle, Guillin Zimmer [6]

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 $t > s > 0$

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

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 (since $w(m_0) = \bigcap_{t\geq 0} \overline{(m_s)_{s\geq t}}$),

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 $V(m_{t_n}) \to V(m^*)$.

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 [16, 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

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

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

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, Siska and S [9].

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

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

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

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

Proof outline: see Chassagneux, S and Tse [3]

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] Carrillo, J. A., McCann, R. J., Villani, C., et al. Kinetic equilibration rates for granular media and related equations: entropy dissipation and mass transportation estimates. *Revista Matematica Iberoamericana 19*, 3 (2003), 971–1018.

[3] Chassagneux, J.-F., Szpruch, L., and Tse, A. Weak quantitative propagation of chaos via differential calculus on the space of measures. *arXiv:1901.02556* (2019).

[4] 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.

[5] 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.

[6] Eberle, A., Guillin, A., and Zimmer, R. Quantitative harris-type theorems for diffusions and mckean–vlasov processes. *Transactions of the American Mathematical Society 371*, 10 (2019), 7135–7173.

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

[8] 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).

References II

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

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

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

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

[13] 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).

[14] 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.

[15] Otto, F. The geometry of dissipative evolution equations: the porous medium equation.

[16] 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.

[17] 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).

References III

[18] Tugaut, J., et al. Convergence to the equilibria for self-stabilizing processes in double-well landscape. *The Annals of Probability 41*, 3A (2013), 1427–1460.

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