From the theory of (stochastic) control to deep learning and back - part 2

Lukasz Szpruch
University of Edinburgh, The Alan Turing Institute, London

Energy functional - Variational Perspective

- We want to minimise

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

where relative entropy H for $m \in \mathcal{P}(\mathbb{R}^d)$

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

and Gibbs measure g:

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

- Mean field Langevin Dynamics

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

- U gives contraction, W smooths the law

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

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

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: Let $f_n : X \to \mathbb{R}$. Recall that f_n Γ-converge to f, if

- for every sequence $x_n \to x$ $f(x) \leq \liminf_{n\to\infty} f_n(x_n)$:
- for every $x \in X$, there is a sequence x_n converging to x such that $f(x) \geq \limsup_{n\to\infty} f_n(x_n)$:

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: Let $f_n : X \to \mathbb{R}$. Recall that f_n Γ-converge to f, if

- for every sequence $x_n \to x$ $f(x) \leq \liminf_{n\to\infty} f_n(x_n)$:
- for every $x \in X$, there is a sequence x_n converging to x such that $f(x) \geq \limsup_{n\to\infty} f_n(x_n)$:

- 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

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. Moreover, $m^\star = \arg\min_{m\in\mathcal{P}(\mathbb{R}^d)} V^\sigma$ iff

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

or equivalently

$$m^\star(x) = \frac{1}{Z} e^{-\frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m^\star, x)} g(x)$$

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

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

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

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

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

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

If $m \notin \mathcal{S}$ then

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

so $m^\star$ is global minimum of V^σ. Since V^σ is strictly convex it is unique.

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

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

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

If $m \notin \mathcal{S}$ then

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

so $m^\star$ is global minimum of V^σ. Since V^σ is strictly convex it is unique.

Step 2 (sufficient condition): Assume m^* satisfies first order condition then for any $\varepsilon > 0$ and $m \in \mathcal{P}(\mathbb{R}^d)$ we have

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

Connection to gradient flow

- Recall

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

Connection to gradient flow

- Recall

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

- If $m^\star$ is such that

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

Connection to gradient flow

- Recall

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

- If $m^\star$ is such that

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

- Then $m^\star$ is a stationary solution of gradient flow PDE

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

Mean-field Langevin equation

We see that if

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

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

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

Mean-field Langevin equation

We see that if

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

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 [Carrillo et al., 2003]
  Otto [Otto, 2001], [Tugaut et al., 2013]
- To establish $\Gamma-$ convergence need result to hold for all σ, so works of [Bogachev et al., 2019] and [Eberle et al., 2019] do not apply.

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.*

Energy Dissipation

Theorem 1

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

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

Energy Dissipation

Theorem 1

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 2

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 MFLD. Then, there exists an invariant measure of of MFLD equal to* $m^\star := \mathrm{argmin}_m V^\sigma(m)$ *and*

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

Convergence

Theorem 2

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 MFLD. Then, there exists an invariant measure of of MFLD equal to* $m^\star := \operatorname{argmin}_m V^\sigma(m)$ *and*

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

If V was continuous then result would follow from tightness of $(m_t)_{t\geq 0}$ and Theorem 1. The entropy is only l.s.c.

Proof key ingredients: Tightness of $(m_t)_{t\geq 0}$, Lasalle's invariance principle, Theorem 1, HWI inequality.

Convergence, step 1: invariance

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

Define ω-limit set

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

Then

i) $\omega(m_0)$ is nonempty and compact (since for any $t \geq 0$, $(m_s)_{s\geq t}$ is relatively compact, $w(m_0) = \bigcap_{t\geq 0} \overline{(m_s)_{s\geq t}}$),
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

Prove that $m^\star \in \omega(m_0)$

Convergence, step 1: invariance

Prove that $m^\star \in \omega(m_0)$

Since $\omega(m_0)$ is compact, there is $\tilde{m} \in \operatorname{argmin}_{m \in \omega(m_0)} V(m)$.

Convergence, step 1: invariance

Prove that $m^\star \in \omega(m_0)$

Since $\omega(m_0)$ is compact, 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 1 for any $s > 0$ we get

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

Convergence, step 1: invariance

Prove that $m^\star \in \omega(m_0)$

Since $\omega(m_0)$ is compact, 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 1 for any $s > 0$ we get

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

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

Convergence, step 1: invariance

Prove that $m^\star \in \omega(m_0)$

Since $\omega(m_0)$ is compact, 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 1 for any $s > 0$ we get

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

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

By Theorem 1

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

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

Convergence, step 2: HWI inequality

$$m^\star \in \omega(m_0) \implies \exists (m_{t_n}) \to m^\star$$

Convergence, step 2: HWI inequality

$m^\star \in \omega(m_0) \implies \exists (m_{t_n}) \to m^\star$

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

Convergence, step 2: HWI inequality

$m^\star \in \omega(m_0) \implies \exists (m_{t_n}) \to m^\star$

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

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 and Villani, 2000]

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

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

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

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

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

Exponential convergence

Theorem 3

If σ is sufficiently large, there exists $\lambda > 0$ s.t

$$\mathcal{W}_2(m_t, m^*) \leq e^{-\lambda t}\mathcal{W}_2(m_0, m^*)\,.$$

Proof see: [Eberle et al., 2019],[Hu et al., 2019]

- New perspective on Lazy training paradigm.

Particle approximation of m^*

Theorem 4

We assume that for any random variables η_1, η_2 such that $\mathbb{E}[|\eta_i|^2] < \infty$, $i = 1, 2$, it holds that

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

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

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

Proof idea [Chassagneux et al., 2019]

Let $(X_i)_{i=1}^N$, $(\tilde{X}_i)_{i=1}^N$ be indepdent i.i.d. with law μ. Let $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ and $m_t^N = \mu + t(\mu_N - \mu)$, $t \in [0,1]$.

Let $(X_i)_{i=1}^N$, $(\tilde{X}_i)_{i=1}^N$ be indepdent i.i.d. with law μ. Let $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ and $m_t^N = \mu + t(\mu_N - \mu)$, $t \in [0,1]$.

By the definition of linear functional derivatives

$$\begin{aligned}\mathbb{E}[F(\mu_N)] - F(\mu) &= \mathbb{E}\left[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_t^N, v)\,(\mu_N - \mu)(\mathrm{d}v)\mathrm{d}t\right] \\ &= \int_0^1 \mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, X_1) - \frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right]\mathrm{d}t\,.\end{aligned}$$

Let $(X_i)_{i=1}^N$, $(\tilde{X}_i)_{i=1}^N$ be indepdent i.i.d. with law μ. Let $\mu_N = \frac{1}{N}\sum_{i=1}^N \delta_{X_i}$ and $m_t^N = \mu + t(\mu_N - \mu)$, $t \in [0,1]$.

By the definition of linear functional derivatives

$$\begin{aligned}\mathbb{E}[F(\mu_N)] - F(\mu) &= \mathbb{E}\left[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m_t^N, v)\,(\mu_N - \mu)(\mathrm{d}v)\mathrm{d}t\right] \\ &= \int_0^1 \mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, X_1) - \frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right]\mathrm{d}t\,.\end{aligned}$$

We introduce the (random) measures

$$\tilde{m}_t^N := m_t^N + \frac{t}{N}(\delta_{\tilde{X}_1} - \delta_{X_1}) \quad \text{and} \quad m_{t,t_1}^N := m_t^N + t_1(\tilde{m}_t^N - m_t^N), \quad t, t_1 \in [0,1],$$

and notice that due to independence of $(X_i)_{i=1}^N$ and $(\tilde{X}_i)_{i=1}^N$ we have that

$$\mathbb{E}\left[\frac{\delta F}{\delta m}(\tilde{m}_t^N, \tilde{X}_1)\right] = \mathbb{E}\left[\frac{\delta F}{\delta m}(m_t^N, X_1)\right].$$

Therefore,

$$\mathbb{E}[F(\mu_N) - F(\mu)] = \int_0^1 \mathbb{E}\left[\frac{\delta F}{\delta m}(\tilde{m}_t^N, \tilde{X}_1) - \frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right] \mathrm{d}t$$

$$= \int_0^1 \mathbb{E}\left[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N, \tilde{X}_1, y_1)(\tilde{m}_t^N - m_t^N)(\mathrm{d}y_1)\,\mathrm{d}t_1\right] \mathrm{d}t$$

$$= \frac{1}{N}\mathbb{E}\left[\int_0^1 \int_0^1 \int_{\mathbb{R}^d} t\frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N, \tilde{X}_1, y_1)(\delta_{\tilde{X}_1} - \delta_{X_1})(\mathrm{d}y_1)\,\mathrm{d}t_1\mathrm{d}t\right] \leq \frac{2L}{N}$$

Therefore,

$$\mathbb{E}[F(\mu_N) - F(\mu)] = \int_0^1 \mathbb{E}\left[\frac{\delta F}{\delta m}(\tilde{m}_t^N, \tilde{X}_1) - \frac{\delta F}{\delta m}(m_t^N, \tilde{X}_1)\right] \mathrm{d}t$$
$$= \int_0^1 \mathbb{E}\left[\int_0^1 \int_{\mathbb{R}^d} \frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N, \tilde{X}_1, y_1)(\tilde{m}_t^N - m_t^N)(\mathrm{d}y_1)\,\mathrm{d}t_1\right] \mathrm{d}t$$
$$= \frac{1}{N}\mathbb{E}\left[\int_0^1 \int_0^1 \int_{\mathbb{R}^d} t\frac{\delta^2 F}{\delta m^2}(m_{t,t_1}^N, \tilde{X}_1, y_1)(\delta_{\tilde{X}_1} - \delta_{X_1})(\mathrm{d}y_1)\,\mathrm{d}t_1\mathrm{d}t\right] \leq \frac{2L}{N}$$

Let $(X_i^*)_{i=1}^N$ be i.i.d. such that $X_i^* \sim m^*$, $i = 1, \ldots, N$. Note that

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

Hence

$$0 \leq \inf_{(x_i)_{i=1}^N \subset \mathbb{R}^d} F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x_i}\right) - F(m^*) \leq \frac{2L}{N}.$$

Mean-Field Neural ODEs

Example

- Recurrent neural networks can be written as

$$X^{l+1} = X^l + h_l \, \phi(X^l, \theta^l) \quad X^0 = \xi \in \mathbb{R}^d$$

Example

- Recurrent neural networks can be written as

$$X^{l+1} = X^l + h_l \, \phi(X^l, \theta^l) \quad X^0 = \xi \in \mathbb{R}^d$$

- Infinitly deep network (useful when fitting time-series data)

$$dX_t^{\xi}(\theta) = \phi(X_t^{\xi}(\theta), \theta_t) \, dt \,, \quad t \in [0, 1] \,, \quad X_0 = \xi \in \mathbb{R}^d.$$

Example

- Recurrent neural networks can be written as

$$X^{l+1} = X^l + h_l \, \phi(X^l, \theta^l) \quad X^0 = \xi \in \mathbb{R}^d$$

- Infinitly deep network (useful when fitting time-series data)

$$dX_t^\xi(\theta) = \phi(X_t^\xi(\theta), \theta_t) \, dt \,, \quad t \in [0,1] \,, \quad X_0 = \xi \in \mathbb{R}^d.$$

- Take input-output data $(\xi, \zeta) \sim \mathcal{M}$. Our objective is to minimize

$$J\Big((\theta)_{t\in[0,T]}\Big) := \int_{\mathbb{R}^d \times \mathbb{R}^d} |\zeta - X_T^\xi(\theta)|^2 \, \mathcal{M}(d\xi, d\zeta) \,.$$

Example

- Recurrent neural networks can be written as

$$X^{l+1} = X^l + h_l \, \phi(X^l, \theta^l) \quad X^0 = \xi \in \mathbb{R}^d$$

- Infinitly deep network (useful when fitting time-series data)

$$dX_t^\xi(\theta) = \phi(X_t^\xi(\theta), \theta_t) \, dt \,, \quad t \in [0,1] \,, \quad X_0 = \xi \in \mathbb{R}^d.$$

- Take input-output data $(\xi, \zeta) \sim \mathcal{M}$. Our objective is to minimize

$$J\Big((\theta)_{t\in[0,T]}\Big) := \int_{\mathbb{R}^d \times \mathbb{R}^d} |\zeta - X_T^\xi(\theta)|^2 \, \mathcal{M}(d\xi, d\zeta) \,.$$

- Compute

$$\frac{d}{d\varepsilon} J\Big(\theta + \varepsilon(\tilde{\theta} - \theta)\Big)\Big|_{\varepsilon=0} = -2 \int_{\mathbb{R}^{2d}} (\zeta - X_T^\xi(\theta)) \frac{d}{d\varepsilon} X_T^\xi(\theta + \varepsilon(\tilde{\theta} - \theta))\Big|_{\varepsilon=0} \mathcal{M}(d\xi, d\zeta) \,.$$

Example

- Recurrent neural networks can be written as

$$X^{l+1} = X^l + h_l\,\phi(X^l, \theta^l) \quad X^0 = \xi \in \mathbb{R}^d$$

- Infinitly deep network (useful when fitting time-series data)

$$dX_t^\xi(\theta) = \phi(X_t^\xi(\theta), \theta_t)\,dt\,, \quad t \in [0,1]\,, \quad X_0 = \xi \in \mathbb{R}^d.$$

- Take input-output data $(\xi, \zeta) \sim \mathcal{M}$. Our objective is to minimize

$$J\Big((\theta)_{t\in[0,T]}\Big) := \int_{\mathbb{R}^d\times\mathbb{R}^d} |\zeta - X_T^\xi(\theta)|^2\,\mathcal{M}(d\xi, d\zeta)\,.$$

- Compute

$$\frac{d}{d\varepsilon} J\Big(\theta+\varepsilon(\tilde\theta-\theta)\Big)\Big|_{\varepsilon=0} = -2\int_{\mathbb{R}^{2d}} (\zeta - X_T^\xi(\theta))\frac{d}{d\varepsilon} X_T^\xi(\theta+\varepsilon(\tilde\theta-\theta))\Big|_{\varepsilon=0}\,\mathcal{M}(d\xi, d\zeta)\,.$$

-

$$\frac{d}{d\varepsilon} X_T^\xi(\theta+\varepsilon(\tilde\theta-\theta))\Big|_{\varepsilon=0} = \int_0^T e^{\int_t^T (\nabla_x\phi)(X_r^\xi(\theta),\theta_r)\,dr} (\nabla_a\phi)(X_t^\xi(\theta),\theta_t)(\tilde\theta_t-\theta_t)\,dt$$

Example

- Let $P_t^{\xi,\zeta}(\theta) := 2(\zeta - X_T^{\xi,\theta}) e^{\int_t^T (\nabla_x \phi)(X_r^{\xi,\cdot}(\theta),\theta_r)\, dr}$ so that

$$dP_t^{\xi,\zeta}(\theta) = -(\nabla_x \phi)(X_t^{\xi}(\theta),\theta_t) P_t^{\xi,\zeta}(\theta)\, dt\,, \;\; P_T^{\xi,\zeta}(\theta) = 2(\zeta - X_T^{\xi,\zeta}(\theta))$$

- Hence $\frac{d}{d\varepsilon} J\Big(\theta + \varepsilon(\tilde{\theta} - \theta)\Big)\Big|_{\varepsilon=0}$ is given by

$$-\int_0^T \int_{\mathbb{R}^{2d}} (\nabla_a \phi)(X_t^{\xi}(\theta),\theta_t)\, P_t^{\xi,\zeta}(\theta)\, \mathcal{M}(d\xi, d\zeta)(\tilde{\theta}_t - \theta_t)\, dt$$

Example

- Let $P_t^{\xi,\zeta}(\theta) := 2(\zeta - X_T^{\xi,\theta}) e^{\int_t^T (\nabla_x \phi)(X_r^{\xi,\theta}(\theta), \theta_r)\, dr}$ so that

  $$dP_t^{\xi,\zeta}(\theta) = -(\nabla_x \phi)(X_t^{\xi}(\theta), \theta_t) P_t^{\xi,\zeta}(\theta)\, dt\,, \;\; P_T^{\xi,\zeta}(\theta) = 2(\zeta - X_T^{\xi,\zeta}(\theta))$$

- Hence $\frac{d}{d\varepsilon} J\Big(\theta + \varepsilon(\tilde{\theta} - \theta)\Big)\Big|_{\varepsilon=0}$ is given by

  $$-\int_0^T \int_{\mathbb{R}^{2d}} (\nabla_a \phi)(X_t^{\xi}(\theta), \theta_t)\, P_t^{\xi,\zeta}(\theta)\, \mathcal{M}(d\xi, d\zeta)(\tilde{\theta}_t - \theta_t)\, dt$$

- This means that for some $\gamma > 0$ choosing

  $$\tilde{\theta}_t = \theta_t + \gamma \int_{\mathbb{R}^{2d}} (\nabla_a \phi)(X_t^{\xi}(\theta), \theta_t) P_t^{\xi,\zeta}(\theta)\, \mathcal{M}(d\xi, d\zeta)$$

  ensures

  $$\frac{d}{d\varepsilon} J\Big(\theta + \varepsilon(\tilde{\theta} - \theta)\Big)\Big|_{\varepsilon=0} \leq 0$$

Relaxed Stochastic Control and Deep Learning

See [Hu et al., 2019, Li et al., 2017, Cuchiero et al., 2019]

- Take external data $\mathcal{D} = (\xi, \zeta) \sim \mathcal{M} \in \mathcal{P}_p(\mathbb{R}^d \times \mathcal{S})$

Relaxed Stochastic Control and Deep Learning

See [Hu et al., 2019, Li et al., 2017, Cuchiero et al., 2019]

- Take external data $\mathcal{D} = (\xi, \zeta) \sim \mathcal{M} \in \mathcal{P}_p(\mathbb{R}^d \times \mathcal{S})$
- $\nu = (\nu_s)_{s \in [0,T]}$, relaxed controls $\nu \in \mathscr{M}([0,T] \times \mathbb{R}^p)$

$$\mathcal{V}_2 := \left\{ \nu : \nu(dt, da) = \nu_t(da)dt,\ \nu_t \in \mathcal{P}_2(\mathbb{R}^p),\ \int_0^T \int |a|^2 \nu_t(da)dt < \infty \right\},$$

Relaxed Stochastic Control and Deep Learning

See [Hu et al., 2019, Li et al., 2017, Cuchiero et al., 2019]

- Take external data $\mathcal{D} = (\xi, \zeta) \sim \mathcal{M} \in \mathcal{P}_p(\mathbb{R}^d \times \mathcal{S})$
- $\nu = (\nu_s)_{s \in [0,T]}$, relaxed controls $\nu \in \mathscr{M}([0,T] \times \mathbb{R}^p)$

$$\mathcal{V}_2 := \left\{ \nu : \nu(dt, da) = \nu_t(da)dt,\ \nu_t \in \mathcal{P}_2(\mathbb{R}^p),\ \int_0^T \int |a|^2 \nu_t(da)dt < \infty \right\},$$

- We consider the following controlled process for $t \in [0, T]$

$$X_t^{\nu,\xi,\zeta} = \xi + \int_0^t \int \phi(X_r^{\nu,\xi,\zeta}, a, \zeta)\, \nu_r(da)\, dr$$

Relaxed Stochastic Control and Deep Learning

See [Hu et al., 2019, Li et al., 2017, Cuchiero et al., 2019]

- Take external data $\mathcal{D} = (\xi, \zeta) \sim \mathcal{M} \in \mathcal{P}_p(\mathbb{R}^d \times \mathcal{S})$
- $\nu = (\nu_s)_{s\in[0,T]}$, relaxed controls $\nu \in \mathscr{M}([0,T] \times \mathbb{R}^p)$

$$\mathcal{V}_2 := \left\{ \nu : \nu(dt, da) = \nu_t(da)dt,\ \nu_t \in \mathcal{P}_2(\mathbb{R}^p),\ \int_0^T \int |a|^2 \nu_t(da)dt < \infty \right\},$$

- We consider the following controlled process for $t \in [0, T]$

$$X_t^{\nu,\xi,\zeta} = \xi + \int_0^t \int \phi(X_r^{\nu,\xi,\zeta}, a, \zeta)\, \nu_r(da)\, dr$$

-

$$\bar{J}^\sigma(\nu, \xi, \zeta) := \int_0^T \int f_t(X_t^{\xi,\zeta}(\nu), a, \zeta)\nu_t(da)dt + g(X_T^{\xi,\zeta}(\nu), \zeta) + \frac{\sigma^2}{2} \int_0^T \mathrm{Ent}(\nu_t)dt\,,$$

$$J^{\sigma,\mathcal{M}}(\nu) := \int_{\mathbb{R}^d\times\mathcal{S}} \bar{J}^\sigma(\nu, \xi, \zeta)\, \mathcal{M}(d\xi, d\zeta)$$

Relaxed Stochastic Control and Deep Learning

- Mean-field perspective on neural networks

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}\varphi(\alpha_{n,i}\cdot z+\rho_{i,n}\cdot\zeta)=\int_{\mathbb{R}^d}\beta\varphi(\alpha\cdot z+\rho\cdot\zeta)\,\nu^n(d\beta,d\alpha,d\rho)\,.$$

Relaxed Stochastic Control and Deep Learning

- Mean-field perspective on neural networks

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}\varphi(\alpha_{n,i}\cdot z+\rho_{i,n}\cdot\zeta)=\int_{\mathbb{R}^d}\beta\varphi(\alpha\cdot z+\rho\cdot\zeta)\,\nu^n(d\beta,d\alpha,d\rho)\,.$$

- Let $\phi(z,a,\zeta)=\beta\varphi(\alpha\cdot z+\rho\cdot\zeta)$, and consider

$$X_t^{\nu,\xi,\zeta}=\xi+\int_0^t\int\phi(X_r^{\nu,\xi,\zeta},a,\zeta)\,\nu_r(da)\,dr$$

Relaxed Stochastic Control and Deep Learning

- The goal is to find, for each $t \in [0, T]$ a vector field flow $(b_{s,t})_{s \geq 0}$ such that the measure flow $(\nu_{s,t})_{s \geq 0}$ given by

$$\partial_s \nu_{s,t} = \mathrm{div}(\nu_{s,t}\, b_{s,t})\,, \quad s \geq 0\,, \ \nu_{0,t} = \nu_t^0 \in \mathcal{P}_2(\mathbb{R}^p)\,,$$

  satisfies that $s \mapsto J^\sigma(\nu_{s,.})$ is decreasing.

Relaxed Stochastic Control and Deep Learning

- The goal is to find, for each $t \in [0, T]$ a vector field flow $(b_{s,t})_{s \geq 0}$ such that the measure flow $(\nu_{s,t})_{s \geq 0}$ given by

$$\partial_s \nu_{s,t} = \mathrm{div}(\nu_{s,t}\, b_{s,t})\,, \quad s \geq 0\,, \ \nu_{0,t} = \nu_t^0 \in \mathcal{P}_2(\mathbb{R}^p)\,,$$

  satisfies that $s \mapsto J^\sigma(\nu_{s,.})$ is decreasing.

- Relaxed Hamiltonian

$$H_t^\sigma(x, p, m, \zeta) := \int h_t(x, p, a, \zeta)\, m(da) + \frac{\sigma^2}{2}\mathrm{Ent}(m)$$
$$h_t(x, p, a, \zeta) := \phi_t(x, a, \zeta)p + f_t(x, a, \zeta)$$

Relaxed Stochastic Control and Deep Learning

- The goal is to find, for each $t \in [0, T]$ a vector field flow $(b_{s,t})_{s \geq 0}$ such that the measure flow $(\nu_{s,t})_{s \geq 0}$ given by

$$\partial_s \nu_{s,t} = \mathrm{div}(\nu_{s,t}\, b_{s,t})\,, \quad s \geq 0\,, \ \nu_{0,t} = \nu_t^0 \in \mathcal{P}_2(\mathbb{R}^p)\,,$$

  satisfies that $s \mapsto J^\sigma(\nu_{s,.})$ is decreasing.

- Relaxed Hamiltonian

$$H_t^\sigma(x, p, m, \zeta) := \int h_t(x, p, a, \zeta)\, m(da) + \frac{\sigma^2}{2}\mathrm{Ent}(m)$$
$$h_t(x, p, a, \zeta) := \phi_t(x, a, \zeta)p + f_t(x, a, \zeta)$$

- The adjoint process $P_t^{\xi,\zeta}(\nu) = (\nabla_x g)(X_t^{\xi,\zeta}(\nu), \zeta)$,

$$dP^{\xi,\zeta}(\nu)_t = -(\nabla_x H_t)(X^{\xi,\zeta}(\nu)_t, P_t^{\nu,\xi,\zeta}(\nu), \nu_t)\, dt$$

Assumptions

Assumption 6

i) $\int_{\mathbb{R}^d \times \mathcal{S}} [|\xi|^2 + |\zeta|^2] \, \mathcal{M}(d\xi, d\zeta) < \infty$.

ii) ϕ, $\nabla_a \phi$, $\nabla_x \phi$, f, $\nabla_a f$, $\nabla_x f$ *and* $\nabla_x g$ *are all Lipschitz continuous in* (x, a), *uniformly in* $(t, \zeta) \in [0, T] \times \mathcal{S}$. *Moreover,* $x \mapsto \nabla_x \phi$, $\nabla_x f$ *and* $\nabla_x g$ *are all continuously differentiable.*

iii)

$$\sup_{t \in [0,T]} \int_{\mathbb{R}^d \times \mathcal{S}} \Big[|\nabla_x g(0, \zeta)|^2 + |\phi_t(0, 0, \zeta)|^2 + |\nabla_x f_t(0, 0, \zeta)|^2 + |\nabla_a f_t(0, 0, \zeta)|^2 \Big] \, \mathcal{M}(d\xi, d\zeta) < \infty.$$

Pontryagin's principle

Theorem 5

Fix $\sigma \geq 0$ and let the Assumption 1 hold. If $\nu \in \mathcal{V}_2$ is (locally) optimal then it must solve the following system:

$$\begin{aligned}
\nu_t &= \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^p)} \int_{\mathbb{R}^d \times \mathcal{S}} H_t^\sigma(X_t^{\xi,\zeta}, P_t^{\xi,\zeta}, \mu, \zeta)\, \mathcal{M}(d\xi, d\zeta)\,, \\
dX_t^{\xi,\zeta} &= \Phi(X_t^{\xi,\zeta}, \nu_t, \zeta)\, dt\,, \quad X_0^{\xi,\zeta} = \xi \in \mathbb{R}^d \\
dP_t^{\xi,\zeta} &= -(\nabla_x H_t)(X_t^{\xi,\zeta}, P_t^{\xi,\zeta}, \nu_t, \zeta)\, dt\,, \quad P_T^{\xi,\zeta} = (\nabla_x g)(X_T^{\xi,\zeta}, \zeta)\,.
\end{aligned}$$

Gradient Flow

Theorem 6

Fix $\sigma \geq 0$ and let the Assumption hold. Let $X^{\xi,\zeta}_{s,\cdot}, P^{\xi,\zeta}_{s,\cdot}$ be the forward and backward processes arising from data (ξ, ζ) with control $\nu_{s,\cdot} \in \mathcal{V}_2$. Then

$$\frac{d}{ds} J^\sigma(\nu_{s,\cdot}) =$$
$$-\int_0^T \int \left(\int_{\mathbb{R}^d \times \mathcal{S}} \left(\nabla_a \frac{\delta H^\sigma_t}{\delta m} \right) (X^{\xi,\zeta}_{s,t}, P^{\xi,\zeta}_{s,t}, \nu_{s,t}, a, \zeta) \mathcal{M}(d\xi, d\zeta) \right) b_{s,t}(a) \nu_{s,t}(da) dt \,.$$

Sketch of the Proof

The goal is to find, for each $t \in [0, T]$ a vector field flow $(b_{s,t})_{s \geq 0}$ such that The measure flow $(\nu_{s,t})_{s \geq 0}$ given by

$$\partial_s \nu_{s,t} = \operatorname{div}(\nu_{s,t}\, b_{s,t})\,, \quad s \geq 0\,, \quad \nu_{0,t} = \nu_t^0 \in \mathcal{P}_2(\mathbb{R}^p)\,,$$

satisfies that $s \mapsto J^\sigma(\nu_{s,\cdot})$ is decreasing.

Sketch of the Proof

The goal is to find, for each $t \in [0, T]$ a vector field flow $(b_{s,t})_{s \geq 0}$ such that The measure flow $(\nu_{s,t})_{s \geq 0}$ given by

$$\partial_s \nu_{s,t} = \operatorname{div}(\nu_{s,t}\, b_{s,t})\,, \quad s \geq 0\,, \quad \nu_{0,t} = \nu_t^0 \in \mathcal{P}_2(\mathbb{R}^p)\,,$$

satisfies that $s \mapsto J^\sigma(\nu_{s,\cdot})$ is decreasing.

Let $V_{s,t} := \frac{d}{ds} X_{s,t}$. One can show that

$$dV_{s,t} = \left[(\nabla_x \Phi)(X_{s,t}, \nu_{s,t}, \zeta) V_{s,t} - \int (\nabla_a \phi_t)(X_{s,t}, \nu_{s,t}, a, \zeta)\, b_{s,t}(a)\, \nu_{s,t}(da) \right] dt\,.$$

Since the equation is affine we can solve it and its solution is

$$V_{s,t} = -\int_0^t \int e^{\int_r^t (\nabla_x \Phi_{\bar{r}})(X_{s,\bar{r}}, \nu_{s,\bar{r}}, \zeta)\, d\bar{r}} (\nabla_a \phi_r)(X_{s,r}, a, \zeta)\, b_{s,r}(a)\, \nu_{s,r}(da)\, dr\,.$$

Sketch of the Proof

Further we can show that

$$\frac{d}{ds}\bar{J}^{\sigma}(\nu_{s,\cdot},\xi,\zeta) = -\int_0^T \int \Bigg[(\nabla_x g)(X_{s,T},\zeta) e^{\int_r^T (\nabla_x \Phi_{\bar{r}})(X_{s,\bar{r}},\nu_{s,\bar{r}},\zeta)\, d\bar{r}} (\nabla_a \phi_r)(X_{s,r},a,\zeta) + \frac{\sigma^2}{2}\left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a)\right)\Bigg] b_{s,t}(a)\, \nu_{s,r}(da)\, dt\,.$$

Sketch of the Proof

Further we can show that

$$\frac{d}{ds}\bar{J}^{\sigma}(\nu_{s,\cdot},\xi,\zeta) = -\int_0^T \int \left[(\nabla_x g)(X_{s,T},\zeta)e^{\int_r^T (\nabla_x \Phi_{\bar{r}})(X_{s,\bar{r}},\nu_{s,\bar{r}},\zeta)\,d\bar{r}}(\nabla_a \phi_r)(X_{s,r},a,\zeta) + \frac{\sigma^2}{2}\left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a)\right)\right] b_{s,t}(a)\,\nu_{s,r}(da)\,dt\,.$$

Now we define

$$P^{\xi,\zeta}_{s,r} := (\nabla_x g)(X_{s,T},\zeta)e^{\int_r^T (\nabla_x \Phi_{\bar{r}})(X_{s,\bar{r}},\nu_{s,\bar{r}},\zeta)\,d\bar{r}}$$

so that

$$\frac{d}{ds}\bar{J}^{\sigma}(\nu_{s,\cdot},\xi,\zeta) = -\int_0^T \int \left[(\nabla_a \phi_r)(X_{s,r},a,\zeta)P_{s,r} + \frac{\sigma^2}{2}\left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a)\right)\right] b_{s,r}(a)\nu_{s,r}(da)\,dr\,.$$

Sketch of the Proof

At this point it is clear how to choose the flow to make this negative: we must take

$$b_{s,r}(a) := \int_{\mathbb{R}^d \times \mathcal{S}} (\nabla_a \phi_r)(X_{s,r}^{\xi,\zeta}, a, \zeta) P_{s,r}^{\xi,\zeta} \, \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a) \right)$$

so that

$$\begin{aligned} &\frac{d}{ds} J^\sigma(\nu_{s,\cdot}) \\ &= -\int_0^T \int \left| \int (\nabla_a \phi_r)(X_{s,r}^{\xi,\zeta}, a, \zeta) P_{s,r}^{\xi,\zeta} \, \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a) \right) \right|^2 \nu_{s,r}(da) dr \\ &\leq 0 \,. \end{aligned}$$

Sketch of the Proof

At this point it is clear how to choose the flow to make this negative: we must take

$$b_{s,r}(a) := \int_{\mathbb{R}^d \times \mathcal{S}} (\nabla_a \phi_r)(X_{s,r}^{\xi,\zeta}, a, \zeta) P_{s,r}^{\xi,\zeta} \, \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a) \right)$$

so that

$$\frac{d}{ds} J^\sigma(\nu_{s,\cdot})$$

$$= - \int_0^T \int \left| \int (\nabla_a \phi_r)(X_{s,r}^{\xi,\zeta}, a, \zeta) P_{s,r}^{\xi,\zeta} \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \left(\frac{\nabla_a \nu_{s,t}(a)}{\nu_{s,t}(a)} + \nabla_a U(a) \right) \right|^2 \nu_{s,r}(da) dr$$

$$\leq 0 \, .$$

with the choice of $b_{s,r}$ made above $\nu_{s,\cdot} = \mathcal{L}(\theta_{s,\cdot})$

$$d\theta_{s,t} = - \left(\int (\nabla_a \phi_t)(X_t^{\xi,\zeta}(\mathcal{L}(\theta_{s,\cdot}), \theta_{s,t}, \zeta) \, P_t^{\xi,\zeta}(\mathcal{L}(\theta_{s,\cdot})) \, \mathcal{M}(d\xi, d\zeta) - \frac{\sigma^2}{2} U(\theta_{s,t}) \right) ds + \sigma \, dB_s$$

Assumption 7 (For existence, uniqueness and invariant measure)

i) *Let* $\int_0^T \mathbb{E}[|\theta_t^0|^q]\, dt < \infty$.

ii) *Let* $\nabla_a U$ *be Lipschitz continuous in* a *and moreover let there* $\kappa > 0$ *such that:*

$$(\nabla_a U(a') - \nabla_a U(a)) \cdot (a' - a) \geq \kappa |a' - a|^2,\ a, a' \in \mathbb{R}^p\,.$$

iii) *Assume that either:*

a) $(x, \zeta) \mapsto \nabla_x g(x, \zeta)$ *is bounded on* $\mathbb{R}^d \times \mathcal{S}$,

b) *or that* $(t, a, \zeta) \mapsto \phi_t(0, a, \zeta)$ *is bounded on* $[0, T] \times \mathbb{R}^p \times \mathcal{S}$ *and* $\mathcal{M}$ *has compact support.*

Theorem 7

Assume that J is bounded from below and that there exists $\bar{\nu}$ such that $J^\sigma(\bar{\nu}) < \infty$ and that $\sigma > 0$. Then

i) $\operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu) \neq \emptyset$,

Theorem 7

Assume that J is bounded from below and that there exists $\bar{\nu}$ such that $J^\sigma(\bar{\nu}) < \infty$ and that $\sigma > 0$. Then

i) $\operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu) \neq \emptyset$,

ii) *if $\nu^\star \in \operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu)$ then for a.a. $t \in (0, T)$ we have that*

$$\mathbf{h}_t(a, \nu^\star, \mathcal{M}) + \frac{\sigma^2}{2} \log(\nu^\star(a)) + \frac{\sigma^2}{2} U(a) \text{ is constant for a.a. } a \in \mathbb{R}^p$$

and $\nu^\star$ is an invariant measure for gradient flow PDE.

Theorem 7

Assume that J is bounded from below and that there exists $\bar{\nu}$ such that $J^\sigma(\bar{\nu}) < \infty$ and that $\sigma > 0$. Then

i) $\operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu) \neq \emptyset$,

ii) *if $\nu^\star \in \operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu)$ then for a.a. $t \in (0, T)$ we have that*

$$\mathbf{h}_t(a, \nu^\star, \mathcal{M}) + \frac{\sigma^2}{2} \log(\nu^\star(a)) + \frac{\sigma^2}{2} U(a) \text{ is constant for a.a. } a \in \mathbb{R}^p$$

and $\nu^\star$ is an invariant measure for gradient flow PDE. Moreover,

iii) *if $\sigma^2 \kappa - 4L > 0$ then $\nu^\star$ is unique.*

Theorem 7

Assume that J is bounded from below and that there exists $\bar{\nu}$ such that $J^\sigma(\bar{\nu}) < \infty$ and that $\sigma > 0$. Then

i) $\operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu) \neq \emptyset$,

ii) *if $\nu^\star \in \operatorname{argmin}_{\nu \in \mathcal{V}_2} J^\sigma(\nu)$ then for a.a. $t \in (0, T)$ we have that*

$$\mathbf{h}_t(a, \nu^\star, \mathcal{M}) + \frac{\sigma^2}{2} \log(\nu^\star(a)) + \frac{\sigma^2}{2} U(a) \text{ is constant for a.a. } a \in \mathbb{R}^p$$

and $\nu^\star$ is an invariant measure for gradient flow PDE. Moreover,

iii) *if $\sigma^2 \kappa - 4L > 0$ then $\nu^\star$ is unique.*

iv) *We have that for all $s \geq 0$ any $\mathcal{L}(\theta_{0,\cdot})$*

$$\mathcal{W}_2^T(\mathcal{L}(\theta_{s,\cdot}), \nu^\star)^2 \leq e^{-\lambda s} \mathcal{W}_2^T(\mathcal{L}(\theta_{0,\cdot}), \nu^\star)^2 .$$

- $$\mathbf{h}_t(a, \mu, \mathcal{M}) := \int_{\mathbb{R}^d \times \mathcal{S}} h_t(X_t^{\xi,\zeta}(\mu), P_t^{\xi,\zeta}(\mu), a, \zeta) \mathcal{M}(d\xi, d\zeta) ,$$

Propagation of chaos

- For $s \in [0, S]$, $t \in [0, T]$ and $1 \leq i \leq N_2$ define

$$\theta^i_{s,t} = \theta^i_{0,t} - \int_0^s \left((\nabla_a \mathbf{h}_t)(\theta^i_{v,t}, \nu^{N_2}_v, \mathcal{M}^{N_1}) + \frac{\sigma^2}{2} (\nabla_a U)(\theta^i_{v,t}) \right) dv + \sigma B^i_s \,,$$

where $\nu^{N_2}_v = \frac{1}{N_2} \sum_{j=1}^{N_2} \delta_{\theta^j_v}$.

Theorem 8

Fix $\lambda = \frac{\sigma^2 \kappa}{2} - \frac{L}{2}(3 + T) + \frac{1}{2}$. Then, there exists c, independent of s, N_1, N_2, p, d, such that, for all i

$$\int_0^T \mathbb{E}\left[\left| \theta^1_{s,t} - \theta^{i,\infty}_{s,t} \right|^2 \right] dt \leq \frac{c}{\lambda}(1 - e^{-\lambda s}) \left(\frac{1}{N_1} + \frac{1}{N_2} \right) .$$

- The rate does not depend on the dimension!

Euler Scheme

$$\widetilde{\theta}^i_{l+1,t} = \widetilde{\theta}^i_{l,t} - \left((\nabla_a \mathbf{h}_t) \left(\widetilde{\theta}^i_{l,t}, \widetilde{\nu}^{N_2}_l, \mathcal{M}^{N_1} \right) + \frac{\sigma^2}{2} U(\widetilde{\theta}^i_{l,t}) \right) (s_{l+1} - s_l) + \sigma (B^i_{s_l+1} - B^i_{s_l}) ,$$

where $\widetilde{\nu}^{N_2}_v = \frac{1}{N_2} \sum_{j_2=1}^{N_2} \delta_{\widetilde{\theta}^{j_2}_v}$.

Theorem 9

There exists $c > 0$ independent of N_1, N_2, p, d, such that, for all $n > 0$,

$$\mathbb{E} \int_0^T |\theta^i_{s_n,t} - \widetilde{\theta}^i_{n,t}|^2 \, dt \leq c \max_{1 \leq l \leq n} |s_l - s_{l-1}|^2 ,$$

provided that $\sigma^2 \kappa$ is large relative to LT.

Generalisation Error

- Recall the cost function

$$J^{\sigma,\mathcal{M}}(\nu) := \int_{\mathbb{R}^d \times \mathcal{S}} \left[\int_0^T \int f_t(X_t^{\nu,\xi,\zeta}, a, \zeta)\, \nu_t(da)\, dt + g(X_T^{\nu,\xi,\zeta}, \zeta) \right] \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \int_0^T \mathrm{Ent}(\nu_t)\, dt\,.$$

Generalisation Error

- Recall the cost function

$$J^{\sigma,\mathcal{M}}(\nu) := \int_{\mathbb{R}^d \times \mathcal{S}} \left[\int_0^T \int f_t(X_t^{\nu,\xi,\zeta}, a, \zeta)\, \nu_t(da)\, dt + g(X_T^{\nu,\xi,\zeta}, \zeta) \right] \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \int_0^T \mathrm{Ent}(\nu_t)\, dt\,.$$

- In practice, one does not have access to population distribution $\mathcal{M}$, but works with finite sample $\mathcal{M}^{N_1} := \frac{1}{N_1} \sum_{j_1}^{N_1} \delta_{(\xi^{j_1}, \zeta^{j_1})}$

Generalisation Error

- Recall the cost function

$$J^{\sigma,\mathcal{M}}(\nu) := \int_{\mathbb{R}^d \times \mathcal{S}} \left[\int_0^T \int f_t(X_t^{\nu,\xi,\zeta}, a, \zeta)\, \nu_t(da)\, dt + g(X_T^{\nu,\xi,\zeta}, \zeta) \right] \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \int_0^T \mathrm{Ent}(\nu_t)\, dt\,.$$

- In practice, one does not have access to population distribution $\mathcal{M}$, but works with finite sample $\mathcal{M}^{N_1} := \frac{1}{N_1} \sum_{j_1}^{N_1} \delta_{(\xi^{j_1}, \zeta^{j_1})}$
- Engineers use $J^{\mathcal{M}^{N_1}}(\nu)$ and NOT $J^{\sigma,\mathcal{M}_{N_1}}(\nu)$ to set stopping criteria for learning

Generalisation Error

- Recall the cost function

$$J^{\sigma,\mathcal{M}}(\nu) := \int_{\mathbb{R}^d \times \mathcal{S}} \left[\int_0^T \int f_t(X_t^{\nu,\xi,\zeta}, a, \zeta)\, \nu_t(da)\, dt + g(X_T^{\nu,\xi,\zeta}, \zeta) \right] \mathcal{M}(d\xi, d\zeta) + \frac{\sigma^2}{2} \int_0^T \mathrm{Ent}(\nu_t)\, dt\,.$$

- In practice, one does not have access to population distribution $\mathcal{M}$, but works with finite sample $\mathcal{M}^{N_1} := \frac{1}{N_1} \sum_{j_1}^{N_1} \delta_{(\xi^{j_1}, \zeta^{j_1})}$
- Engineers use $J^{\mathcal{M}^{N_1}}(\nu)$ and NOT $J^{\sigma,\mathcal{M}_{N_1}}(\nu)$ to set stopping criteria for learning
- The entropy term can be viewed as implicit regularisation

Theorem 10

Assume that $\sigma^2\kappa$ is sufficiently. Then there is $c > 0$ independent of λ, S, N_1, N_2, d, p s.t

$$\mathbb{E}\left[\left|J^{\mathcal{M}}(\nu^{\star,\sigma}) - J^{\mathcal{M}}(\nu_{S,\cdot}^{\sigma,N_1,N_2,\Delta s})\right|^2\right] \leq c\left(e^{-\lambda S} + \frac{1}{N_1} + \frac{1}{N_2} + h\right),$$

where $h := \max_{0<s_l<S}(s_l - s_{l-1})$. The generalisation error is given by

$$J^{\mathcal{M}}(\nu_{S,\cdot}^{\sigma,N_1,N_2,\Delta s})$$
$$= J^{\mathcal{M}}(\nu_{S,\cdot}^{\sigma,N_1,N_2,\Delta s}) - J^{\mathcal{M}}(\nu^{\star,\sigma}) - \frac{\sigma^2}{2}\int_0^T Ent(\nu^{\star,\sigma}) + \min_{\mu\in\mathcal{V}_2} J^{\sigma,\mathcal{M}}(\mu),$$

since $\min_{\mu\in\mathcal{V}_2} J^{\sigma,\mathcal{M}}(\mu) = J^{\sigma,\mathcal{M}}(\nu^{*,\sigma})$.

- N_1 - size of the training data
- N_2 - proxy to the the number of parameters
- γ - learning rate; S/γ - proxy for training time
- By discretising ODEs we can get estimateson the number of layers

Assumption 8

Fix $\epsilon > 0$ *and* $N_1 > 0$. *Assume that* $\forall \mathcal{M}^{N_1}$, $J^{M^{N_1}}\left(\nu^{\star,\sigma,N_1}\right) \leq \epsilon$.

Theorem 11

There is $c > 0$ *independent of* λ, S, N_1, N_2, d, p *s.t*

$$\mathbb{E}\left[\left|J^{\mathcal{M}}\left(\nu_{S,\cdot}^{\sigma,N_1,N_2,\Delta s}\right)\right|^2\right] \leq \epsilon^2 + c\left(e^{-\lambda S} + \frac{1}{N_1} + \frac{1}{N_2} + h\right).$$

References I

[Bogachev et al., 2019] Bogachev, V., Röckner, M., and Shaposhnikov, S. (2019). On convergence to stationary distributions for solutions of nonlinear fokker–planck–kolmogorov equations. *Journal of Mathematical Sciences*, 242(1):69–84.

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

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

[Cuchiero et al., 2019] Cuchiero, C., Larsson, M., and Teichmann, J. (2019). Deep neural networks, generic universal interpolation, and controlled odes. *arXiv preprint arXiv:1908.07838*.

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

[Hu et al., 2019] Hu, K., Kazeykina, A., and Ren, Z. (2019). Mean-field langevin system, optimal control and deep neural networks. *arXiv preprint arXiv:1909.07278*.

[Li et al., 2017] Li, Q., Chen, L., Tai, C., and Weinan, E. (2017). Maximum principle based algorithms for deep learning. *The Journal of Machine Learning Research*, 18(1):5998–6026.

[Otto, 2001] Otto, F. (2001). The geometry of dissipative evolution equations: the porous medium equation.

References II

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

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