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

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

Figure: David Siska

Figure: Zhenjie Ren

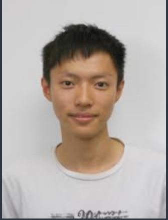


Figure: Kaitong Hu

Figure: Jean-Francois Jabir

Outline

- Sampling vs optimisation on $\mathbb{R}^d$
- Mean-Field Langevin Dynamics - training of one hidden layer neural network viewed as an optimisation problem over Wassersatin space, [Hu et al., 2019b].
- Neural ODEs via Relaxed Optimal Control. A perspective on deep recurrent neural networks, [Jabir et al., 2019].
- Gradient flows for (regularised) stochastic control problem, [Šiška and Szpruch, 2020].
- Robust pricing and hedging with neural SDEs, [Gierjatowicz et al., 2020].
- Unbiased approximation of parametric path dependent PDEs, [Vidales et al., 2018].

Key messages of this mini course

- Training neural nets is a sampling problem

Key messages of this mini course

- Training neural nets is a sampling problem
- Gradient flow view on training neural networks provides mathematical framework to study machine learning

Key messages of this mini course

- Training neural nets is a sampling problem
- Gradient flow view on training neural networks provides mathematical framework to study machine learning
- Probabilistic numerical analysis provides quantitative bounds that do not suffer from the curse of dimensionality

Key messages of this mini course

- Training neural nets is a sampling problem
- Gradient flow view on training neural networks provides mathematical framework to study machine learning
- Probabilistic numerical analysis provides quantitative bounds that do not suffer from the curse of dimensionality
- Machine learning perspective leads to new algorithms and mathematical tools for (stochastic) control problems and offers a fresh perspective on classical quantitative finance problems.

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)
  - vastly larger data sets

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)
  - vastly larger data sets
  - Benchmark data sets (MNIST, ImageNet, CIFAR)

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)
  - vastly larger data sets
  - Benchmark data sets (MNIST, ImageNet, CIFAR)
  - GPUs, TPUs and cloud computing

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)
  - vastly larger data sets
  - Benchmark data sets (MNIST, ImageNet, CIFAR)
  - GPUs, TPUs and cloud computing
  - very efficient open source libraries (Tensorflow, Theano, Torch).

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)
  - vastly larger data sets
  - Benchmark data sets (MNIST, ImageNet, CIFAR)
  - GPUs, TPUs and cloud computing
  - very efficient open source libraries (Tensorflow, Theano, Torch).
- "Imagenet classification with deep convolutional neural networks" by Krizhevsky, Sutskever, Hinton, (2012) NIPS - 68613 citation as of 31-08-2020.

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
  - architectural innovations (e.g. many layers, dropout, LSTMs)
  - algorithmic innovations (e.g ADAM methods)
  - vastly larger data sets
  - Benchmark data sets (MNIST, ImageNet, CIFAR)
  - GPUs, TPUs and cloud computing
  - very efficient open source libraries (Tensorflow, Theano, Torch).
- "Imagenet classification with deep convolutional neural networks" by Krizhevsky, Sutskever, Hinton, (2012) NIPS - 68613 citation as of 31-08-2020.
- "Grandmaster level in StarCraft II using multi-agent reinforcement learning" by Vinyals, Babuschkin,...,David Silver, (2019) Nature. Estimated cost of training the algorithm - $30m.

Success of Deep Learning

- Neural networks appeared in the 1943 seminal work by Warren McCulloch and Walter Pitts inspired by certain functionalities of the brain aiming for artificial intelligence (AI)
- Excellent performance (image and language recognition, classification tasks, etc) due to
    - architectural innovations (e.g. many layers, dropout, LSTMs)
    - algorithmic innovations (e.g ADAM methods)
    - vastly larger data sets
    - Benchmark data sets (MNIST, ImageNet, CIFAR)
    - GPUs, TPUs and cloud computing
    - very efficient open source libraries (Tensorflow, Theano, Torch).
- "Imagenet classification with deep convolutional neural networks" by Krizhevsky, Sutskever, Hinton, (2012) NIPS - 68613 citation as of 31-08-2020.
- "Grandmaster level in StarCraft II using multi-agent reinforcement learning" by Vinyals, Babuschkin,...,David Silver, (2019) Nature. Estimated cost of training the algorithm - $30m.
- So far deep learning is successful in a 'relatively' stationary regime.

Neural networks

i) an *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$; $\varphi(z) = (\varphi(z_1), \ldots, \varphi(z_l))^\top$

Neural networks

i) an *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$; $\varphi(z) = (\varphi(z_1), \dots, \varphi(z_l))^\top$

ii) The space of parameters

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

Neural networks

i) an *activation function* $\varphi : \mathbb{R} \to \mathbb{R}$; $\varphi(z) = (\varphi(z_1), \ldots, \varphi(z_l))^\top$

ii) The space of parameters

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

iii) 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

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

Arnold-Kolmogorov theorem and Universal approximation

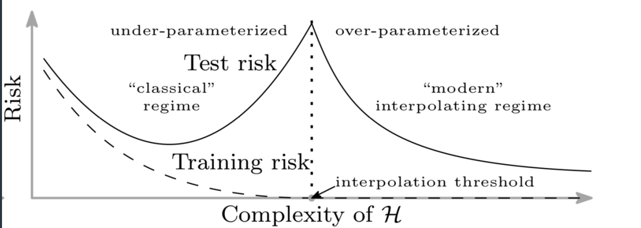
If an activation function φ is bounded, continuous and non-constant, then for any compact set $K \subset \mathbb{R}^d$ the set

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

is dense in the space of continuous functions from K to $\mathbb{R}$. See e.g. Hornik [Hornik, 1991],[Cybenko, 1989].

- Practical quantitative results are possible when working with additional structural assumptions (e.g low-dimensional hypothesis)
- Some recent work [Schmidt-Hieber et al., 2020], [Ma et al., 2019]

New era of overparameterized statistical models ?



From Belkin. et.al. [Belkin et al., 2018].

- Need for new theory to study generalisation error. Classical Vapnik dimension and Rademacher complexity doesn't help.
- Overparametrised models can be optimal in the high signal-to-noise ratio regime Montanari et.al [Mei and Montanari, 2019]
- Implicit Regularisation [Heiss et al., 2019], [Neyshabur et al., 2017]

Key Questions

i) **Function approximation theory**: the challenge is to derive non-asymptotic results; expressiveness in terms of width and depth; network architecture design: feed-forward, convolutional, LSTM, ResNet, Attention Networks...

ii) **Generalisation error** in particular in overparametrised regime.

iii) **Non-convex optimisation and effect of noise in stochastic gradient algorithms**, in general non-convex optimisation problems are NP-hard; links with the optimisation; lazy and mean-field regimes in overparametrised setting

(Noisy) Gradient Descent

Optimisation on $\mathbb{R}^d$

- Consider $F : \mathbb{R}^d \to \mathbb{R}$

Optimisation on $\mathbb{R}^d$

- Consider $F : \mathbb{R}^d \to \mathbb{R}$
- Assume F is strongly convex with parameter $m > 0$ i.e $\forall x, y$ and $\alpha \in (0, 1)$ we have
  - $$(x - y, \nabla_x F(x) - \nabla_x F(y)) \geq m\,|x - y|^2$$

Optimisation on $\mathbb{R}^d$

- Consider $F : \mathbb{R}^d \to \mathbb{R}$
- Assume F is strongly convex with parameter $m > 0$ i.e $\forall x, y$ and $\alpha \in (0,1)$ we have
  - $$(x - y, \nabla_x F(x) - \nabla_x F(y)) \geq m\,|x - y|^2$$
  - equivalently
    $$F(y) \geq F(x) + (\nabla_x F(x), y - x) + \frac{m}{2}|y - x|^2$$

Optimisation on $\mathbb{R}^d$

- Consider $F : \mathbb{R}^d \to \mathbb{R}$
- Assume F is strongly convex with parameter $m > 0$ i.e $\forall x, y$ and $\alpha \in (0, 1)$ we have
  - $$(x - y, \nabla_x F(x) - \nabla_x F(y)) \geq m\, |x - y|^2$$
  - equivalently
    $$F(y) \geq F(x) + (\nabla_x F(x), y - x) + \frac{m}{2}|y - x|^2$$
  - equivalently
    $$F(\alpha + (1 - \alpha)y) \leq \alpha F(x) + (1 - \alpha)F(y) - \frac{1}{2}m\alpha(1 - \alpha)|x - y|^2$$

Gradient flow on $\mathbb{R}^d$

- Classical gradient descent algorithm with fixed learning rata $\gamma > 0$

$$x_{n+1} = x_n - \gamma(\nabla_x F)(x_n)$$

Gradient flow on $\mathbb{R}^d$

- Classical gradient descent algorithm with fixed learning rata $\gamma > 0$

$$x_{n+1} = x_n - \gamma(\nabla_x F)(x_n)$$

- Continuous view point aka gradient flow

$$dx_t = -(\nabla_x F)(x_t)dt$$

Gradient flow on $\mathbb{R}^d$

- Classical gradient descent algorithm with fixed learning rata $\gamma > 0$

$$x_{n+1} = x_n - \gamma(\nabla_x F)(x_n)$$

- Continuous view point aka gradient flow

$$dx_t = -(\nabla_x F)(x_t)dt$$

- F is decreasing along gradient flow (x_t)

$$dF(x_t) = (\nabla_x F)(x_t)dx_t = -|(\nabla_x F)(x_t)|^2 dt\,.$$

Gradient flow on $\mathbb{R}^d$

- Classical gradient descent algorithm with fixed learning rata $\gamma > 0$

$$x_{n+1} = x_n - \gamma(\nabla_x F)(x_n)$$

- Continuous view point aka gradient flow

$$dx_t = -(\nabla_x F)(x_t)dt$$

- F is decreasing along gradient flow (x_t)

$$dF(x_t) = (\nabla_x F)(x_t)dx_t = -|(\nabla_x F)(x_t)|^2 dt\,.$$

- Since F is strongly convex $\exists! x^\star$ s.t $F(x^\star) = min_x F(x)$.

Rate of convergence

- We can easily compute the rate of convergence

$$d(x_t - x^*) = -((\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt$$

Rate of convergence

- We can easily compute the rate of convergence

$$d(x_t - x^*) = -((\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt$$

- Hence

$$d|x_t - x^*|^2 = -2(x_t - x^\star, (\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt \leq -2m|x_t - x^\star|$$

Rate of convergence

- We can easily compute the rate of convergence

$$d(x_t - x^*) = -((\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt$$

- Hence

$$d|x_t - x^*|^2 = -2(x_t - x^\star, (\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt \leq -2m|x_t - x^\star|$$

- We have

$$|x_t - x^\star|^2 \leq |x_0 - x^\star|^2 e^{-2mt}$$

Rate of convergence

- We can easily compute the rate of convergence

$$d(x_t - x^*) = -((\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt$$

- Hence

$$d|x_t - x^*|^2 = -2(x_t - x^\star, (\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt \leq -2m|x_t - x^\star|$$

- We have

$$|x_t - x^\star|^2 \leq |x_0 - x^\star|^2 e^{-2mt}$$

- Exercise: Do a computation directly for discrete time dynamics. Need F Lipschitz and the Lipschitz constant matters.

Rate of convergence

- We can easily compute the rate of convergence

$$d(x_t - x^*) = -((\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt$$

- Hence

$$d|x_t - x^*|^2 = -2(x_t - x^\star, (\nabla_x F)(x_t) - (\nabla_x F)(x^\star))dt \leq -2m|x_t - x^\star|$$

- We have

$$|x_t - x^\star|^2 \leq |x_0 - x^\star|^2 e^{-2mt}$$

- Exercise: Do a computation directly for discrete time dynamics. Need F Lipschitz and the Lipschitz constant matters.
- If we drop the assumption that F is convex, gradient descent can only converge to a local minimum.

Noisy gradient descent $\mathbb{R}^d$

- Consider noisy gradient descent with $\sigma > 0$

$$dX_t = -(\nabla_x F)(X_t)dt + \sigma dW_t$$

Noisy gradient descent $\mathbb{R}^d$

- Consider noisy gradient descent with $\sigma > 0$

$$dX_t = -(\nabla_x F)(X_t)dt + \sigma dW_t$$

- A natural question: $\mu_t := \mathcal{L}(X_t) \to ?$ when $t \to \infty$.

Noisy gradient descent $\mathbb{R}^d$

- Consider noisy gradient descent with $\sigma > 0$

$$dX_t = -(\nabla_x F)(X_t)dt + \sigma dW_t$$

- A natural question: $\mu_t := \mathcal{L}(X_t) \to ?$ when $t \to \infty$.
- PDE for the law. Let $\phi \in C^2(\mathbb{R}^d)$

$$\frac{d}{dt}\mathbb{E}[\phi(X_t)] = \mathbb{E}\left[-(\nabla F)(X_t) \cdot \nabla\phi(X_t) + \frac{\sigma^2}{2}\nabla^2\phi(X_t)\right].$$

Noisy gradient descent $\mathbb{R}^d$

- Consider noisy gradient descent with $\sigma > 0$

$$dX_t = -(\nabla_x F)(X_t)dt + \sigma dW_t$$

- A natural question: $\mu_t := \mathcal{L}(X_t) \to ?$ when $t \to \infty$.
- PDE for the law. Let $\phi \in C^2(\mathbb{R}^d)$

$$\frac{d}{dt}\mathbb{E}[\phi(X_t)] = \mathbb{E}\left[-(\nabla F)(X_t)\cdot\nabla\phi(X_t) + \frac{\sigma^2}{2}\nabla^2\phi(X_t)\right].$$

- Suppose that μ_t admits density $\mu(t,x)$

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^d}\phi(x)\mu(t,x)dx &= \int_{\mathbb{R}^d}\left(-(\nabla F)(x)\nabla\phi(x) + \frac{\sigma^2}{2}\nabla^2\phi(x)\right)\mu(t,x)dx \\ &= \int_{\mathbb{R}^d}\left(\mathrm{div}((\nabla F)(x)\mu(t,x)) + \frac{\sigma^2}{2}\nabla^2\mu(t,x)\right)\phi(x)dx\end{aligned}$$

Noisy gradient descent $\mathbb{R}^d$

- Consider noisy gradient descent with $\sigma > 0$

$$dX_t = -(\nabla_x F)(X_t)dt + \sigma dW_t$$

- A natural question: $\mu_t := \mathcal{L}(X_t) \to ?$ when $t \to \infty$.
- PDE for the law. Let $\phi \in C^2(\mathbb{R}^d)$

$$\frac{d}{dt}\mathbb{E}[\phi(X_t)] = \mathbb{E}\left[-(\nabla F)(X_t)\cdot\nabla\phi(X_t) + \frac{\sigma^2}{2}\nabla^2\phi(X_t)\right].$$

- Suppose that μ_t admits density $\mu(t,x)$

$$\begin{aligned}\frac{d}{dt}\int_{\mathbb{R}^d}\phi(x)\mu(t,x)dx &= \int_{\mathbb{R}^d}\left(-(\nabla F)(x)\nabla\phi(x) + \frac{\sigma^2}{2}\nabla^2\phi(x)\right)\mu(t,x)dx \\ &= \int_{\mathbb{R}^d}\left(\mathrm{div}((\nabla F)(x)\mu(t,x)) + \frac{\sigma^2}{2}\nabla^2\mu(t,x)\right)\phi(x)dx\end{aligned}$$

- Since this holds for all ϕ, $\mu = \mu(t,x)$ solves

$$\partial_t\mu = \mathrm{div}((\nabla F)\mu) + \frac{\sigma^2}{2}\Delta\mu$$

Gibbs measure

- Under mild conditions on ∇F, X is ergodic with invariant measure

$$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2}F(x)} dx \quad Z = \int_{\mathbb{R}^d} e^{-\frac{-2}{\sigma^2}F(x)} dx$$

Gibbs measure

- Under mild conditions on ∇F, X is ergodic with invariant measure

$$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2} F(x)} dx \quad Z = \int_{\mathbb{R}^d} e^{-\frac{-2}{\sigma^2} F(x)} dx$$

- In other words for all X_0, $\mu_t = \mathcal{L}(X_t)$ converges weakly to π

Gibbs measure

- Under mild conditions on ∇F, X is ergodic with invariant measure

$$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2} F(x)} dx \quad Z = \int_{\mathbb{R}^d} e^{-\frac{-2}{\sigma^2} F(x)} dx$$

- In other words for all X_0, $\mu_t = \mathcal{L}(X_t)$ converges weakly to π
- Indeed plugging in π into right-hand side of the PDE:

$$\begin{aligned}
&\frac{1}{Z} \int_{\mathbb{R}^d} \left(-\nabla F(x) \nabla \phi(x) + \frac{\sigma^2}{2} \nabla^2 \phi(x) \right) e^{-\frac{2}{\sigma^2} F(x)} dx \\
&= \frac{1}{Z} \int_{\mathbb{R}^d} \left(-\nabla F(x) \nabla \phi(x) + \frac{\sigma^2}{2} \nabla \phi(x) \frac{2}{\sigma^2} \nabla F(x) \right) e^{-\frac{2}{\sigma^2} F(x)} dx = 0 \\
&\implies \frac{d}{dt} \mathbb{E}[\phi(X_t)] = 0
\end{aligned}$$

Gibbs measure

- Under mild conditions on ∇F, X is ergodic with invariant measure

$$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2}F(x)} dx \quad Z = \int_{\mathbb{R}^d} e^{-\frac{-2}{\sigma^2}F(x)} dx$$

- In other words for all X_0, $\mu_t = \mathcal{L}(X_t)$ converges weakly to π
- Indeed plugging in π into right-hand side of the PDE:

$$\begin{aligned}
&\frac{1}{Z}\int_{\mathbb{R}^d} \left(-\nabla F(x)\nabla\phi(x) + \frac{\sigma^2}{2}\nabla^2\phi(x)\right) e^{-\frac{2}{\sigma^2}F(x)} dx \\
&= \frac{1}{Z}\int_{\mathbb{R}^d} \left(-\nabla F(x)\nabla\phi(x) + \frac{\sigma^2}{2}\nabla\phi(x)\frac{2}{\sigma^2}\nabla F(x)\right) e^{-\frac{2}{\sigma^2}F(x)} dx = 0 \\
&\implies \frac{d}{dt}\mathbb{E}[\phi(X_t)] = 0
\end{aligned}$$

- Hence π is a stationary solution to the PDE. Extra work needed to prove that $\mu_t \Rightarrow \pi$.

Laplace method

- $$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2} F(x)} dx$$

Laplace method

- $$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2} F(x)} dx$$

- Consider $\delta > 0$

$$\pi(F(X) > \min F + \delta) = \frac{1}{Z} \int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx$$
$$\leq \frac{\int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx}{\int 1_{\{F(x) \leq \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx}$$

Laplace method

- $$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2} F(x)} dx$$

- Consider $\delta > 0$

$$\pi(F(X) > \min F + \delta) = \frac{1}{Z} \int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx$$

$$\leq \frac{\int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx}{\int 1_{\{F(x) \leq \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx}$$

- $F(x) \leq \min F + \delta \implies \frac{1}{e^{-F(x)}} \leq \frac{1}{e^{-(\min F + \delta)}}$

$$\pi(F(x) > \min F + \delta) \leq \frac{\int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2}(F(x) - (\min F + \delta))} dx}{\int 1_{\{F(x) \leq \min F + \delta\}} dx} \to 0 \text{ as } \sigma \to 0$$

Laplace method

- $$\pi(dx) = \frac{1}{Z} e^{-\frac{2}{\sigma^2} F(x)} dx$$

- Consider $\delta > 0$

$$\pi(F(X) > \min F + \delta) = \frac{1}{Z} \int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx$$
$$\leq \frac{\int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx}{\int 1_{\{F(x) \leq \min F + \delta\}} e^{-\frac{2}{\sigma^2} F(x)} dx}$$

- $F(x) \leq \min F + \delta \implies \frac{1}{e^{-F(x)}} \leq \frac{1}{e^{-(\min F + \delta)}}$

$$\pi(F(x) > \min F + \delta) \leq \frac{\int 1_{\{F(x) > \min F + \delta\}} e^{-\frac{2}{\sigma^2}(F(x) - (\min F + \delta))} dx}{\int 1_{\{F(x) \leq \min F + \delta\}} dx} \to 0 \text{ as } \sigma \to 0$$

- As $\sigma \to 0$ the π concentrates near minimiser of F
- No Convexity required!. See [Hwang, 1980].

Differential Calculus on $\mathcal{P}(\mathbb{R}^d)$

Measure derivatives

Definition 1 (functional/flat derivative or first variation)

We say that $V : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is $\mathcal{C}^1$ if there exists a continuous map $\frac{\delta V}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that for any $m, m' \in \mathcal{P}(\mathbb{R}^d)$

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

- Note $\frac{\delta V}{\delta m}$ is defined up to normalising constant. We take

$$\int_{\mathbb{R}^d} \frac{\delta V}{\delta m}(m,y)m(dy) = 0$$

Measure derivatives

Definition 1 (functional/flat derivative or first variation)

We say that $V : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is $\mathcal{C}^1$ if there exists a continuous map $\frac{\delta V}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that for any $m, m' \in \mathcal{P}(\mathbb{R}^d)$

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

- Note $\frac{\delta V}{\delta m}$ is defined up to normalising constant. We take

$$\int_{\mathbb{R}^d} \frac{\delta V}{\delta m}(m,y)m(dy) = 0$$

- Take $\lambda \in (0,1)$. Define $m^\lambda := m + \lambda(m'-m)$ and note that

$$V(m') - V(m) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta V}{\delta m}(m^\lambda, y)(m'-m)(dy)d\lambda$$

Measure derivatives

Definition 1 (functional/flat derivative or first variation)

We say that $V : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ *is* $\mathcal{C}^1$ *if there exists a continuous map* $\frac{\delta V}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ *such that for any* $m, m' \in \mathcal{P}(\mathbb{R}^d)$

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

- Note $\frac{\delta V}{\delta m}$ is defined up to normalising constant. We take

$$\int_{\mathbb{R}^d} \frac{\delta V}{\delta m}(m,y)m(dy) = 0$$

- Take $\lambda \in (0,1)$. Define $m^\lambda := m + \lambda(m'-m)$ and note that

$$V(m') - V(m) = \int_0^1 \int_{\mathbb{R}^d} \frac{\delta V}{\delta m}(m^\lambda, y)(m'-m)(dy)d\lambda$$

- Note that regularity of $\frac{\delta V}{\delta m}(m,y)$ in y may determine the metric (e.g total variation or Wasserstein) in which V is Lipschitz.

Intrinsic/Lions/Wasserstein derivative

Definition 2

If $\frac{\delta V}{\delta m}$ *is* C^1 *in* y *the intrinsic derivative* $D_m V : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}^d$ *is defined by*

$$D_m V(m, y) := \left(\nabla_y \frac{\delta V}{\delta m} \right)(m, y)$$

Lemma 1 ([Cardaliaguet et al., 2015])

Assume that V *is* $\mathcal{C}^1$ *with* $\frac{\delta V}{\delta m}$ *is* C^1 *in* y *and* $D_m V$ *is continuous in both variables. Let* $b : \mathbb{R}^d \to \mathbb{R}^d$ *be a Borel measurable and bounded. Then*

$$\lim_{s \searrow 0} \frac{V((Id + sb)\#m) - V(m)}{s} = \int_{\mathbb{R}^d} D_m V(m)(y) \cdot b(y) m(dy) .$$

Intrinsic/Lions/Wasserstein derivative

Proof.

Let $m^{s,\lambda} := m + \lambda((Id + sb)\#m - m)$. Then by change of variables formula and mean value theorem

$$V((Id + sb)\#m) - V(m) = \int_0^1 \int \frac{\delta V}{\delta m}(m^{s,\lambda}, y)((Id + s\, b)\#m - m)(dy) d\lambda$$

$$= \int_0^1 \int \left(\frac{\delta V}{\delta m}(m^{s,\lambda}, y + s\, b(y)) - \frac{\delta V}{\delta m}(m^{s,\lambda}, y) \right) m(dy) d\lambda$$

$$= s \int_0^1 \int \int_0^1 D_m V(m^{s,\lambda}, y + t\, s\, b(y)) b(y) dt\, m(dy) d\lambda$$

$\blacksquare$

- Example: $V(m) = \int_{\mathbb{R}^d} f(x)\, m(dx) = (f, m)$.

$$\frac{\delta V}{\delta m}(m, y) = f(y) \text{ and } \quad D_m V(m, y) = \nabla_y f(y)\,.$$

Chain Rule - proof due to Paul-Eric Chaudru de Raynal

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \quad X_0 \in L^2$$

- Need for the chain rule on $\mathcal{P}_2(\mathbb{R}^d)$ to compute $\frac{d}{dt} V(\mathcal{L}(X_t))$

Recall $\mu^{\lambda,\epsilon} = \mu_t + \lambda(\mu_{t+\epsilon} - \mu_t)$, $\mu^{\lambda,\epsilon} \to \mu_t$ when $\epsilon \to 0$.

$$\frac{d}{dt} V(\mu_t) = \lim_{\epsilon \searrow 0} \epsilon^{-1}(V(\mu_{t+\epsilon}) - V(\mu_t))$$

Chain Rule - proof due to Paul-Eric Chaudru de Raynal

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \quad X_0 \in L^2$$

- Need for the chain rule on $\mathcal{P}_2(\mathbb{R}^d)$ to compute $\frac{d}{dt} V(\mathcal{L}(X_t))$

Recall $\mu^{\lambda,\epsilon} = \mu_t + \lambda(\mu_{t+\epsilon} - \mu_t)$, $\mu^{\lambda,\epsilon} \to \mu_t$ when $\epsilon \to 0$.

$$\frac{d}{dt} V(\mu_t) = \lim_{\epsilon \searrow 0} \epsilon^{-1} (V(\mu_{t+\epsilon}) - V(\mu_t))$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int_0^1 \int \frac{\delta V}{\delta m}(\mu_t^{\lambda,\epsilon}, y)(\mu_{t+\epsilon} - \mu_t)(dy) d\lambda$$

Chain Rule - proof due to Paul-Eric Chaudru de Raynal

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \quad X_0 \in L^2$$

- Need for the chain rule on $\mathcal{P}_2(\mathbb{R}^d)$ to compute $\frac{d}{dt}V(\mathcal{L}(X_t))$

Recall $\mu^{\lambda,\epsilon} = \mu_t + \lambda(\mu_{t+\epsilon} - \mu_t)$, $\mu^{\lambda,\epsilon} \to \mu_t$ when $\epsilon \to 0$.

$$\frac{d}{dt}V(\mu_t) = \lim_{\epsilon \searrow 0} \epsilon^{-1}(V(\mu_{t+\epsilon}) - V(\mu_t))$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int_0^1 \int \frac{\delta V}{\delta m}(\mu_t^{\lambda,\epsilon}, y)(\mu_{t+\epsilon} - \mu_t)(dy) d\lambda$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int \mathbb{E}\left[\frac{\delta V}{\delta m}(\mu_t, X_{t+\epsilon}^{t,y}) - \frac{\delta V}{\delta m}(\mu_t, y)\right] \mu_t(dy)$$

Chain Rule - proof due to Paul-Eric Chaudru de Raynal

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \quad X_0 \in L^2$$

- Need for the chain rule on $\mathcal{P}_2(\mathbb{R}^d)$ to compute $\frac{d}{dt} V(\mathcal{L}(X_t))$

Recall $\mu^{\lambda,\epsilon} = \mu_t + \lambda(\mu_{t+\epsilon} - \mu_t)$, $\mu^{\lambda,\epsilon} \to \mu_t$ when $\epsilon \to 0$.

$$\frac{d}{dt} V(\mu_t) = \lim_{\epsilon \searrow 0} \epsilon^{-1}(V(\mu_{t+\epsilon}) - V(\mu_t))$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int_0^1 \int \frac{\delta V}{\delta m}(\mu_t^{\lambda,\epsilon}, y)(\mu_{t+\epsilon} - \mu_t)(dy) d\lambda$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int \mathbb{E}\left[\frac{\delta V}{\delta m}(\mu_t, X_{t+\epsilon}^{t,y}) - \frac{\delta V}{\delta m}(\mu_t, y)\right] \mu_t(dy)$$

$$= \int \mathbb{E}^{t,y}\left[\frac{d}{ds}\left(\frac{\delta V}{\delta m}(\mu_t, X_s^{t,y})\right) |_{s=t}\right] \mu_t(dy)$$

Chain Rule - proof due to Paul-Eric Chaudru de Raynal

$$X_t = X_0 + \int_0^t b_s ds + \int_0^t \sigma_s dW_s, \quad X_0 \in L^2$$

- Need for the chain rule on $\mathcal{P}_2(\mathbb{R}^d)$ to compute $\frac{d}{dt} V(\mathcal{L}(X_t))$

Recall $\mu^{\lambda,\epsilon} = \mu_t + \lambda(\mu_{t+\epsilon} - \mu_t)$, $\mu^{\lambda,\epsilon} \to \mu_t$ when $\epsilon \to 0$.

$$\frac{d}{dt} V(\mu_t) = \lim_{\epsilon \searrow 0} \epsilon^{-1} (V(\mu_{t+\epsilon}) - V(\mu_t))$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int_0^1 \int \frac{\delta V}{\delta m}(\mu_t^{\lambda,\epsilon}, y)(\mu_{t+\epsilon} - \mu_t)(dy) d\lambda$$

$$= \lim_{\epsilon \searrow 0} \epsilon^{-1} \int \mathbb{E}\left[\frac{\delta V}{\delta m}(\mu_t, X_{t+\epsilon}^{t,y}) - \frac{\delta V}{\delta m}(\mu_t, y)\right] \mu_t(dy)$$

$$= \int \mathbb{E}^{t,y}\left[\frac{d}{ds}\left(\frac{\delta V}{\delta m}(\mu_t, X_s^{t,y})\right) |_{s=t}\right] \mu_t(dy)$$

$$= \int \left[b_t D_m V(\mu_t, y) + \frac{1}{2}(\sigma\sigma^T)_t \nabla_y D_m(\mu_t, y)\right] \mu_t(dy)$$

Variational perspective on noisy gradient descent

Variational perspective

- Define
$$V^\sigma(m) := \int F(x)m(dx) + \frac{\sigma^2}{2}H(m),$$
where relative entropy H for $m \in \mathcal{P}(\mathbb{R}^d)$
$$H(m) := \begin{cases} \int_{\mathbb{R}^d} m(x)\log m(x)dx & \text{if } m \text{ is a.c. w.r.t. Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$

Variational perspective

- Define
$$V^\sigma(m) := \int F(x)m(dx) + \frac{\sigma^2}{2}H(m),$$
where relative entropy H for $m \in \mathcal{P}(\mathbb{R}^d)$
$$H(m) := \begin{cases} \int_{\mathbb{R}^d} m(x) \log m(x) dx & \text{if } m \text{ is a.c. w.r.t. Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$

Variational perspective

- Define
$$V^\sigma(m) := \int F(x)m(dx) + \frac{\sigma^2}{2}H(m),$$
where relative entropy H for $m \in \mathcal{P}(\mathbb{R}^d)$
$$H(m) := \begin{cases} \int_{\mathbb{R}^d} m(x)\log m(x)dx & \text{if } m \text{ is a.c. w.r.t. Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$
- Let $b : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a vector field and consider gradient flow (we take b so that PDE is well defined)
$$\partial_t \nu_t = \mathrm{div}(b_t \nu_t)$$

Variational perspective

- Define
$$V^\sigma(m) := \int F(x)m(dx) + \frac{\sigma^2}{2}H(m),$$
where relative entropy H for $m \in \mathcal{P}(\mathbb{R}^d)$
$$H(m) := \begin{cases} \int_{\mathbb{R}^d} m(x)\log m(x)dx & \text{if } m \text{ is a.c. w.r.t. Lebesgue measure} \\ \infty & \text{otherwise} \end{cases}$$
- Let $b : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ be a vector field and consider gradient flow (we take b so that PDE is well defined)
$$\partial_t \nu_t = \mathrm{div}(b_t \nu_t)$$
- Find b so that $V^\sigma(\nu_t) \searrow$ as $t \to \infty$

Variational perspective

- For $\epsilon, \lambda > 0$ let $\nu_t^{\lambda,\epsilon} := \nu_t + \lambda(\nu_{t+\epsilon} - \nu_t)$ we have

$$
\begin{aligned}
\partial_t V^\sigma(\nu_t) &= \lim_{\epsilon \to 0} \epsilon^{-1} \left(V^\sigma(\nu_{t+\epsilon}) - V^\sigma(\nu_t) \right) \\
&= \lim_{\epsilon \to 0} \epsilon^{-1} \left(\int_0^1 \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t^{\lambda,\epsilon}, y)(\nu_{t+\epsilon} - \nu_t)(dy) d\lambda \right)
\end{aligned}
$$

Variational perspective

- For $\epsilon, \lambda > 0$ let $\nu_t^{\lambda,\epsilon} := \nu_t + \lambda(\nu_{t+\epsilon} - \nu_t)$ we have

$$\begin{aligned}\partial_t V^\sigma(\nu_t) &= \lim_{\epsilon \to 0} \epsilon^{-1} \left(V^\sigma(\nu_{t+\epsilon}) - V^\sigma(\nu_t)\right) \\ &= \lim_{\epsilon \to 0} \epsilon^{-1} \left(\int_0^1 \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t^{\lambda,\epsilon}, y)(\nu_{t+\epsilon} - \nu_t)(dy) d\lambda \right)\end{aligned}$$

- Note that $\nu_t^{\lambda,\epsilon} \to \nu_t$ as $\epsilon \to 0$ hence

$$\begin{aligned}\partial_t V^\sigma(\nu_t) &= \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t, y)\partial_t \nu_t(dy) = \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t, y)\mathrm{div}(b_t \nu_t)(dy) \\ &= -\int \left(\nabla_y \frac{\delta V^\sigma}{\delta \nu}\right)(\nu_t, y) b_t \nu_t(dy)\end{aligned}$$

Variational perspective

- For $\epsilon, \lambda > 0$ let $\nu_t^{\lambda,\epsilon} := \nu_t + \lambda(\nu_{t+\epsilon} - \nu_t)$ we have

$$\begin{aligned}\partial_t V^\sigma(\nu_t) &= \lim_{\epsilon\to 0} \epsilon^{-1}\left(V^\sigma(\nu_{t+\epsilon}) - V^\sigma(\nu_t)\right)\\ &= \lim_{\epsilon\to 0} \epsilon^{-1}\left(\int_0^1 \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t^{\lambda,\epsilon}, y)(\nu_{t+\epsilon} - \nu_t)(dy)d\lambda\right)\end{aligned}$$

- Note that $\nu_t^{\lambda,\epsilon} \to \nu_t$ as $\epsilon \to 0$ hence

$$\begin{aligned}\partial_t V^\sigma(\nu_t) &= \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t, y)\partial_t \nu_t(dy) = \int \frac{\delta V^\sigma}{\delta \nu}(\nu_t, y)\mathrm{div}(b_t \nu_t)(dy)\\ &= -\int \left(\nabla_y \frac{\delta V^\sigma}{\delta \nu}\right)(\nu_t, y) b_t \nu_t(dy)\end{aligned}$$

- To have $V^\sigma(\nu_t) \searrow$ take

$$b_t(y) := \left(\nabla_y \frac{\delta V^\sigma}{\delta \nu}\right)(\nu_t, y)$$

Variational perspective

- Recall that $V^\sigma(m) = (F, m) + \frac{\sigma^2}{2}(\log m, m)$

$$\frac{\delta V^\sigma}{\delta m}(m, y) = F(y) + \frac{\sigma^2}{2}(\log m(y) + 1)$$

$$b_t(y) = \left(\nabla_y \frac{\delta V^\sigma}{\delta m}\right)(m, y) = (\nabla_y F)(y) + \frac{\sigma^2}{2}\nabla_y \log(m(y))$$

Variational perspective

- Recall that $V^\sigma(m) = (F, m) + \frac{\sigma^2}{2}(\log m, m)$

$$\frac{\delta V^\sigma}{\delta m}(m, y) = F(y) + \frac{\sigma^2}{2}(\log m(y) + 1)$$

$$b_t(y) = \left(\nabla_y \frac{\delta V^\sigma}{\delta m}\right)(m, y) = (\nabla_y F)(y) + \frac{\sigma^2}{2}\nabla_y \log(m(y))$$

- Plug this back into the gradient flow equation

$$\partial_t \nu_t = \mathrm{div}\left(\left((\nabla F) + \frac{\sigma^2}{2}\nabla \log(\nu_t)\right)\nu_t\right)$$

$$\partial_t \nu_t = \mathrm{div}\left((\nabla F)\nu_t\right) + \frac{\sigma^2}{2}\Delta \nu_t$$

Variational perspective

- Recall that $V^\sigma(m) = (F, m) + \frac{\sigma^2}{2}(\log m, m)$

$$\frac{\delta V^\sigma}{\delta m}(m, y) = F(y) + \frac{\sigma^2}{2}(\log m(y) + 1)$$

$$b_t(y) = \left(\nabla_y \frac{\delta V^\sigma}{\delta m}\right)(m, y) = (\nabla_y F)(y) + \frac{\sigma^2}{2}\nabla_y \log(m(y))$$

- Plug this back into the gradient flow equation

$$\partial_t \nu_t = \mathrm{div}\left(\left((\nabla F) + \frac{\sigma^2}{2}\nabla \log(\nu_t)\right)\nu_t\right)$$

$$\partial_t \nu_t = \mathrm{div}\left((\nabla F)\nu_t\right) + \frac{\sigma^2}{2}\Delta \nu_t$$

- What is a minimiser of V^σ? Note V^σ is strictly convex hence the first order condition

$$\frac{\delta V^\sigma}{\delta m}(m, y) = F(y) + \frac{\sigma^2}{2}(\log m(y) + 1) = const$$

Variational perspective

- Recall that $V^\sigma(m) = (F, m) + \frac{\sigma^2}{2}(\log m, m)$

$$\frac{\delta V^\sigma}{\delta m}(m,y) = F(y) + \frac{\sigma^2}{2}(\log m(y) + 1)$$

$$b_t(y) = \left(\nabla_y \frac{\delta V^\sigma}{\delta m}\right)(m,y) = (\nabla_y F)(y) + \frac{\sigma^2}{2}\nabla_y \log(m(y))$$

- Plug this back into the gradient flow equation

$$\partial_t \nu_t = \mathrm{div}\left(\left((\nabla F) + \frac{\sigma^2}{2}\nabla \log(\nu_t)\right)\nu_t\right)$$

$$\partial_t \nu_t = \mathrm{div}\left((\nabla F)\nu_t\right) + \frac{\sigma^2}{2}\Delta \nu_t$$

- What is a minimiser of V^σ? Note V^σ is strictly convex hence the first order condition

$$\frac{\delta V^\sigma}{\delta m}(m,y) = F(y) + \frac{\sigma^2}{2}(\log m(y) + 1) = const$$

- $$m(y) = e^{-\frac{2}{\sigma^2}F(y)} \cdot const$$

One hidden layer neural network

Non-covex minimization problem

- Consider network

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}\varphi(\alpha_{n,i}\cdot z)=\int_{\mathbb{R}^d}\beta\varphi(\alpha\cdot z)\,m^n(\mathrm{d}\beta,\mathrm{d}\alpha)\,.$$

Non-covex minimization problem

- Consider network

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}\varphi(\alpha_{n,i}\cdot z)=\int_{\mathbb{R}^d}\beta\varphi(\alpha\cdot z)\,m^n(\mathrm{d}\beta,\mathrm{d}\alpha)\,.$$

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

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

which is non-convex.

Non-covex minimization problem

- Consider network

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}\varphi(\alpha_{n,i}\cdot z) = \int_{\mathbb{R}^d}\beta\varphi(\alpha\cdot z)\, m^n(\mathrm{d}\beta,\mathrm{d}\alpha)\,.$$

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

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

which is non-convex.

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

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

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

Non-covex minimization problem

- Consider network

$$\frac{1}{n}\sum_{i=1}^{n}\beta_{n,i}\varphi(\alpha_{n,i}\cdot z)=\int_{\mathbb{R}^d}\beta\varphi(\alpha\cdot z)\,m^n(\mathrm{d}\beta,\mathrm{d}\alpha)\,.$$

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

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

which is non-convex.

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

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

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

- No hope for deterministic gradient to find global minimum....

Approximation with gradient descent

- In practice noisy (regularised), gradient descent algorithms are used:

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

where ξ_k^i are i.i.d. samples from $N(0, I_d)$.

Approximation with gradient descent

- In practice noisy (regularised), gradient descent algorithms are used:

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

  where ξ^i_k are i.i.d. samples from $N(0, I_d)$.
- Taking weak limit gives

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

Mean-field limit and convexity

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

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

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

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 [Mei et al., 2018], Chizat and Bach [Chizat and Bach, 2018] as well as Rotskoff and Vanden-Eijnden [Rotskoff and Vanden-Eijnden, 2018].

Derivation of MFLD

- $$F^N(x^1, \dots, x^N) = F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x^i}\right) = \int_{\mathbb{R}^d} \Phi\left(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x^j, z)\right) \nu(\mathrm{d}z, \mathrm{d}y).$$
- Then
$$\mathrm{d}X_t^i = -\left(N\partial_{x_i} F^N(X_t^1, \dots, X_t^N) + \frac{\sigma^2}{2}\nabla U(X_t^i)\right)\mathrm{d}t + \sigma \mathrm{d}W_t^i\,.$$

Derivation of MFLD

- $$F^N(x^1,\dots,x^N) = F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x^i}\right) = \int_{\mathbb{R}^d} \Phi\left(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x^j, z)\right) \nu(\mathrm{d}z, \mathrm{d}y).$$
- Then
$$\mathrm{d}X_t^i = -\left(N\partial_{x_i} F^N(X_t^1,\dots,X_t^N) + \frac{\sigma^2}{2}\nabla U(X_t^i)\right)\mathrm{d}t + \sigma \mathrm{d}W_t^i\,.$$
- We expect to have, as $N \to \infty$,
$$\begin{cases} dX_t = -\left(D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla U(X_t)\right) dt + \sigma dW_t \quad t \in [0,\infty) \\ m_t = \mathrm{Law}(X_t) \quad t \in [0,\infty)\,. \end{cases}$$

Derivation of MFLD

- $$F^N(x^1,\dots,x^N) = F\left(\frac{1}{N}\sum_{i=1}^N \delta_{x^i}\right) = \int_{\mathbb{R}^d} \Phi\left(y - \frac{1}{N}\sum_{j=1}^N \hat{\varphi}(x^j,z)\right)\nu(\mathrm{d}z,\mathrm{d}y).$$
- Then
$$\mathrm{d}X_t^i = -\left(N\partial_{x_i}F^N(X_t^1,\dots,X_t^N) + \frac{\sigma^2}{2}\nabla U(X_t^i)\right)\mathrm{d}t + \sigma \mathrm{d}W_t^i\,.$$
- We expect to have, as $N\to\infty$,
$$\begin{cases} dX_t = -\left(D_mF(m_t,X_t) + \frac{\sigma^2}{2}\nabla U(X_t)\right)\,dt + \sigma dW_t \quad t\in[0,\infty) \\ m_t = \mathrm{Law}(X_t) \quad t\in[0,\infty)\,. \end{cases}$$
- Fokker–Planck
$$\partial_t m = \nabla\cdot\left(\left(D_mF(m,\cdot) + \frac{\sigma^2}{2}\nabla U\right)m + \frac{\sigma^2}{2}\nabla m\right) \text{ on } (0,\infty)\times\mathbb{R}^d\,.$$

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 3

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

Assumption 4

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 5

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 5

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 5

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 6

Under Assumption 3 and 4, 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

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

or equivalently

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

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

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

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. $V(\bar{m}) < \infty$,

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

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^* is global minimum of V. Since V 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. $V(\bar{m}) < \infty$,

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

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^* is global minimum of V. Since V 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

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

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

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

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

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

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

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

Mean-field Langevin equation

We see that if

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

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 7

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

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

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

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

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

Theorem 2

Let $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumption 4 and 7, 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 2

Let $m_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Under Assumption 4 and 7, 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 3

Let Assumption 3, 4 and 7 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* $m^* := \operatorname{argmin}_m V^\sigma(m)$ *and*

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

Convergence

Theorem 3

Let Assumption 3, 4 and 7 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 $m^* := \operatorname{argmin}_m V^\sigma(m)$ and

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

If V was continuous then result would follow from tightness of $(m_t)_{t\geq 0}$ and Theorem 2. The entropy is only l.s.c.

Proof key ingredients: Tightness of $(m_t)_{t\geq 0}$, Lasalle's invariance principle, Theorem 2, 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 2 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 2 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 2 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 2

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

Due to the first order condition (Proposition 6) 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

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

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 4

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., 2019a]

- New perspective on Lazy training paradigm.

References I

[Belkin et al., 2018] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2018). Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118.*

[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.

[Cardaliaguet et al., 2015] Cardaliaguet, P., Delarue, F., Lasry, J.-M., and Lions, P.-L. (2015). The master equation and the convergence problem in mean field games. *arXiv preprint arXiv:1509.02505.*

[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.

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

[Cybenko, 1989] Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314.

[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.

[Gierjatowicz et al., 2020] Gierjatowicz, P., Sabate-Vidales, M., Siska, D., Szpruch, L., and Zuric, Z. (2020). Robust pricing and hedging via neural sdes. *Available at SSRN 3646241.*

[Heiss et al., 2019] Heiss, J., Teichmann, J., and Wutte, H. (2019). How implicit regularization of neural networks affects the learned function–part i. *arXiv preprint arXiv:1911.02903.*

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

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

References II

[Hu et al., 2019b] Hu, K., Ren, Z., Siska, D., and Szpruch, L. (2019b). Mean-field langevin dynamics and energy landscape of neural networks. *arXiv preprint arXiv:1905.07769.*

[Hwang, 1980] Hwang, C.-R. (1980). Laplace's method revisited: weak convergence of probability measures. *The Annals of Probability*, pages 1177–1182.

[Jabir et al., 2019] Jabir, J.-F., Šiška, D., and Szpruch, Ł. (2019). Mean-field neural odes via relaxed optimal control. *arXiv preprint arXiv:1912.05475.*

[Ma et al., 2019] Ma, C., Wu, L., et al. (2019). Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039.*

[Mei and Montanari, 2019] Mei, S. and Montanari, A. (2019). The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355.*

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

[Neyshabur et al., 2017] Neyshabur, B., Tomioka, R., Salakhutdinov, R., and Srebro, N. (2017). Geometry of optimization and implicit regularization in deep learning. *arXiv preprint arXiv:1705.03071.*

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

[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.

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

[Schmidt-Hieber et al., 2020] Schmidt-Hieber, J. et al. (2020). Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897.

[Šiška and Szpruch, 2020] Šiška, D. and Szpruch, Ł. (2020). Gradient flows for regularized stochastic control problems. *arXiv preprint arXiv:2006.05956.*

References III

[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.

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