Mean-Field Langevin Dynamics and Neural Networks

Zhenjie Ren

CEREMADE, Université Paris-Dauphine

joint works with Giovanni Conforti, Kaitong Hu, Anna Kazeykina, David Siska, Lukasz Szpruch, Xiaolu Tan, Junjian Yang

MATH-IMS Joint Applied Mathematics Colloquium Series

August 28, 2020

Dauphine | PSL UNIVERSITÉ PARIS CEREMADE UMR CNRS 7534

Classical Langevin dynamics and non-convex optimization

The Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x f$ subject to damping and random collision.

Overdamped Langevin dynamics

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

Underdamped Langevin dynamics

$$dX_t = V_t dt$$
$$dV_t = \big(-\nabla_x f(X_t) - \gamma V_t\big)dt + \sigma dW_t$$

Classical Langevin dynamics and non-convex optimization

The Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x f$ subject to damping and random collision.

Overdamped Langevin dynamics

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

Underdamped Langevin dynamics

$$dX_t = V_t dt$$
$$dV_t = \big(-\nabla_x f(X_t) - \gamma V_t\big)dt + \sigma dW_t$$

Under mild conditions, the two Markov diffusions admits unique invariant measures whose densities read:

Overdamped Langevin dynamics

$$m^*(x) = Ce^{-\frac{2}{\sigma^2}f(x)}$$

Underdamped Langevin dynamics

$$m^*(x, v) = Ce^{-\frac{2}{\sigma^2}\left(f(x)+\frac{1}{2}|v|^2\right)}$$

Classical Langevin dynamics and non-convex optimization

The Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x f$ subject to damping and random collision.

Overdamped Langevin dynamics

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

Underdamped Langevin dynamics

$$dX_t = V_t dt$$
$$dV_t = \big(-\nabla_x f(X_t) - \gamma V_t\big)dt + \sigma dW_t$$

Under mild conditions, the two Markov diffusions admits unique invariant measures whose densities read:

Overdamped Langevin dynamics

$$m^*(x) = Ce^{-\frac{2}{\sigma^2}f(x)} \to \delta_{\arg\min f(x)}$$

Underdamped Langevin dynamics

$$m^*(x, v) = Ce^{-\frac{2}{\sigma^2}\left(f(x)+\frac{1}{2}|v|^2\right)}$$

Classical Langevin dynamics and non-convex optimization

The Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x f$ subject to damping and random collision.

Overdamped Langevin dynamics

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

Underdamped Langevin dynamics

$$dX_t = V_t dt$$
$$dV_t = \big(-\nabla_x f(X_t) - \gamma V_t\big)dt + \sigma dW_t$$

Under mild conditions, the two Markov diffusions admits unique invariant measures whose densities read:

Overdamped Langevin dynamics

$$m^*(x) = Ce^{-\frac{2}{\sigma^2}f(x)} \to \delta_{\arg\min f(x)}$$

Underdamped Langevin dynamics

$$m^*(x,v) \to \delta_{\arg\min f(x) + \frac{1}{2}|v|^2}$$

Classical Langevin dynamics and non-convex optimization

The Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x f$ subject to damping and random collision.

Overdamped Langevin dynamics

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

Underdamped Langevin dynamics

$$dX_t = V_t dt$$
$$dV_t = \big(-\nabla_x f(X_t) - \gamma V_t\big)dt + \sigma dW_t$$

Under mild conditions, the two Markov diffusions admits unique invariant measures whose densities read:

Overdamped Langevin dynamics

$$m^*(x) = Ce^{-\frac{2}{\sigma^2}f(x)} \to \delta_{\arg\min f(x)}$$

Underdamped Langevin dynamics

$$m^*(x, v) \to \delta_{(\arg\min f(x), 0)}$$

Classical Langevin dynamics and non-convex optimization

The Langevin dynamics was first introduced in statistical physics to describe the motion of a particle with position X and velocity V in a potential field $\nabla_x f$ subject to damping and random collision.

Overdamped Langevin dynamics

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

Underdamped Langevin dynamics

$$dX_t = V_t dt$$
$$dV_t = \big(-\nabla_x f(X_t) - \gamma V_t\big)dt + \sigma dW_t$$

Under mild conditions, the two Markov diffusions admits unique invariant measures whose densities read:

Overdamped Langevin dynamics

$$m^*(x) = Ce^{-\frac{2}{\sigma^2}f(x)} \to \delta_{\arg\min f(x)}$$

Underdamped Langevin dynamics

$$m^*(x, v) \to \delta_{(\arg\min f(x), 0)}$$

In particular, f does NOT need to be convex.

Relation with classical algorithms

If we overlook the Brownian noise, then

- the overdamped process $\Rightarrow$ gradient descent algorithm
- the underdamped process $\Rightarrow$ Hamiltonian gradient descent algorithm

But their convergence to the minimizer is ensured only for convex potential function f.

Relation with classical algorithms

If we overlook the Brownian noise, then

- the overdamped process $\Rightarrow$ gradient descent algorithm
- the underdamped process $\Rightarrow$ Hamiltonian gradient descent algorithm

But their convergence to the minimizer is ensured only for convex potential function f.

Taking into account the Brownian noise with constant σ, we may produce samplings of the invariant measures

- the overdamped Langevin $\Leftrightarrow$ MCMC
- the underdamped Langevin $\Leftrightarrow$ Hamiltonian MCMC

The convergence rate of MCMC algorithm is in general dimension dependent !

Relation with classical algorithms

If we overlook the Brownian noise, then

- the overdamped process $\Rightarrow$ gradient descent algorithm
- the underdamped process $\Rightarrow$ Hamiltonian gradient descent algorithm

But their convergence to the minimizer is ensured only for convex potential function f.

Taking into account the Brownian noise with constant σ, we may produce samplings of the invariant measures

- the overdamped Langevin $\Leftrightarrow$ MCMC
- the underdamped Langevin $\Leftrightarrow$ Hamiltonian MCMC

The convergence rate of MCMC algorithm is in general dimension dependent !

One may diminish $\sigma \downarrow 0$ along the simulation $\Rightarrow$ Simulation annealing.

Deep neural networks

The deep neural networks have won and continue gaining impressive success in various applications. Mathematically speaking, we may approximate a given function f with the parametrized function:

$$f(z) \approx \varphi_n \circ \cdots \circ \varphi_1(z), \quad \text{where} \quad \varphi_i(z) := \sum_{k=1}^{n_i} c_k^i \varphi(A_k^i z + b_k^i)$$

and φ is a given non-constant, bounded, continuous activation function. The expressiveness of the neural network is ensured by the universal representation theorem. However, the efficiency of such over-parametrized, non-convex optimization is still a mystery for mathematical analysis.

Deep neural networks

The deep neural networks have won and continue gaining impressive success in various applications. Mathematically speaking, we may approximate a given function f with the parametrized function:

$$f(z) \approx \varphi_n \circ \cdots \circ \varphi_1(z), \quad \text{where} \quad \varphi_i(z) := \sum_{k=1}^{n_i} c_k^i \varphi(A_k^i z + b_k^i)$$

and φ is a given non-constant, bounded, continuous activation function. The expressiveness of the neural network is ensured by the universal representation theorem. However, the efficiency of such over-parametrized, non-convex optimization is still a mystery for mathematical analysis.

It is natural to study this problem using Mean-field Langevin equations.

Table of Contents

Two-layer neural network

In the work with *K. Hu, D. Siska, L. Szpruch* '19, we focused on the two-layer network, and aimed at minimizing

$$\inf_{n,(c_k,A_k,b_k)} \mathbf{E}\Big[\Big|f(Z) - \sum_{k=1}^{n} c_k\varphi(A_k Z + b_k)\Big|^2\Big],$$

where Z represents the data and $\mathbf{E}$ is the expectation under the law of the data.

Two-layer neural network

In the work with *K. Hu, D. Siska, L. Szpruch* '19, we focused on the two-layer network, and aimed at minimizing

$$\inf_{n,(c_k,A_k,b_k)} \mathbf{E}\Big[\Big|f(Z) - \frac{1}{n}\sum_{k=1}^{n} c_k\varphi(A_k Z + b_k)\Big|^2\Big],$$

where Z represents the data and $\mathbf{E}$ is the expectation under the law of the data.

Two-layer neural network

In the work with *K. Hu, D. Siska, L. Szpruch* '19, we focused on the two-layer network, and aimed at minimizing

$$\inf_{\nu = Law(C,A,B)} \mathbf{E}\Big[\Big|f(Z) - \mathbb{E}^{\nu}\big[C\varphi(AZ+B)\big]\Big|^2\Big],$$

where Z represents the data and $\mathbf{E}$ is the expectation under the law of the data.

Two-layer neural network

In the work with *K. Hu, D. Siska, L. Szpruch* '19, we focused on the two-layer network, and aimed at minimizing

$$\inf_{\nu = Law(C,A,B)} F(\nu), \quad \text{where} \quad F(\nu) := \mathbf{E}\Big[\big|f(Z) - \mathbb{E}^{\nu}\big[C\varphi(AZ+B)\big]\big|^2\Big],$$

where Z represents the data and $\mathbf{E}$ is the expectation under the law of the data. Note that F is convex in ν.

Two-layer neural network

In the work with *K. Hu, D. Siska, L. Szpruch* '19, we focused on the two-layer network, and aimed at minimizing

$$\inf_{\nu} F(\nu) + \frac{\sigma^2}{2}\mathbf{Ent}(\nu),$$

Note that F is convex in ν. Take $\mathbf{Ent}(\cdot)$, the relative entropy w.r.t. Lebesgue measure, as a regularizer, and note that $\mathbf{Ent}(\cdot)$ is strictly convex.

Two-layer neural network

In the work with *K. Hu, D. Siska, L. Szpruch* '19, we focused on the two-layer network, and aimed at minimizing

$$\inf_{\nu} F(\nu) + \frac{\sigma^2}{2}\mathbf{Ent}(\nu),$$

Note that F is convex in ν. Take $\mathbf{Ent}(\cdot)$, the relative entropy w.r.t. Lebesgue measure, as a regularizer, and note that $\mathbf{Ent}(\cdot)$ is strictly convex.

How to characterize the <u>minimizer</u> of a function of probabilities ?

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^\lambda, x)(m' - m)(dx)d\lambda$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have
  $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big(m^\lambda, x\big)(m' - m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m,x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m,x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^\lambda, x)(m' - m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m,x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m,x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

If further assume F is convex, we have

$$\varepsilon F(m') + (1-\varepsilon)F(m) \;\geq\; F(m^\varepsilon)$$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^\lambda, x)(m' - m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m,x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m,x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

If further assume F is convex, we have

$$\varepsilon\Big(F(m') - F(m)\Big) \;\geq\; F(m^\varepsilon) - F(m)$$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big(m^\lambda, x\big)(m'-m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m,x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m,x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

If further assume F is convex, we have

$$F(m') - F(m) \;\geq\; \frac{1}{\varepsilon}\int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big(m^\lambda, x\big)(m'-m)(dx)d\lambda$$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}\big(m^\lambda, x\big)(m' - m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m, x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m, x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

If further assume F is convex, we have

$$F(m') - F(m) \;\geq\; \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m, x)(m' - m)(dx)$$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^\lambda, x)(m' - m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m,x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m,x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

If further assume F is convex, we have

$$F(m') - F(m) \ \geq \ \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m,x)(m' - m)(dx)$$

Therefore, a sufficient condition for m being a minimizer would be

$$\frac{\delta F}{\delta m}(m,x) = C \quad \text{for all } x$$

Derivatives of functions of probabilities

Let $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$. Denote its derivative by $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$.

- given m, m', denote $m^\lambda := (1-\lambda)m + \lambda m'$ we have $F(m^\varepsilon) - F(m) = \int_0^\varepsilon \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m^\lambda, x)(m'-m)(dx)d\lambda$
- e.g. (a) $F(m) := \mathbb{E}^m[\varphi(X)]$, then $\frac{\delta F}{\delta m}(m,x) = \varphi(x)$
  (b) $F(m) := g\big(\mathbb{E}^m[\varphi(X)]\big)$, then $\frac{\delta F}{\delta m}(m,x) = \dot{g}\big(\mathbb{E}^m[\varphi(X)]\big)\varphi(x)$

If further assume F is convex, we have

$$F(m') - F(m) \;\geq\; \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(m,x)(m'-m)(dx)$$

Therefore, a sufficient condition for m being a minimizer would be

Intrinsic derivative $\quad D_m F(m,x) = \nabla \frac{\delta F}{\delta m}(m,x) = 0 \quad$ for all x

First order condition of minimizers

Under the presence of the entropy regularizer, we can also prove the first order equation is a necessary condition for being minimizer.

Theorem (Hu, R., Siska, Szpruch, '19)

Under mild conditions, if $m^* = \arg\min_m \left\{ F(m) + \frac{\sigma^2}{2} \mathbf{Ent}(m) \right\}$, *then*

$$D_m F(m^*, x) + \frac{\sigma^2}{2} \nabla \ln m^*(x) = 0, \quad \textit{for all } x. \tag{1}$$

Conversely, if F *to be convex,* (1) *implies* m^* *is the minimizer.*

First order condition of minimizers

Under the presence of the entropy regularizer, we can also prove the first order equation is a necessary condition for being minimizer.

Theorem (Hu, R., Siska, Szpruch, '19)

Under mild conditions, if $m^* = \arg\min_m \left\{ F(m) + \frac{\sigma^2}{2} \mathbf{Ent}(m) \right\}$, *then*

$$D_m F(m^*, x) + \frac{\sigma^2}{2} \nabla \ln m^*(x) = 0, \quad \textit{for all } x. \tag{1}$$

Conversely, if F *to be convex,* (1) *implies* m^* *is the minimizer.*

Note that the density of m^* satisfies:

$$m^*(x) = Ce^{-\frac{2}{\sigma^2} \frac{\delta F}{\delta m}(m^*, x)}$$

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

$$D_m F(m^*, x) + \frac{\sigma^2}{2} \nabla \ln m^*(x) = 0$$

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

$$D_m F(m^*, x) + \frac{\sigma^2}{2} \frac{\nabla m^*(x)}{m^*(x)} = 0$$

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

$$D_m F(m^*, x) m^*(x) + \frac{\sigma^2}{2} \nabla m^*(x) = 0$$

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

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

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

$$\nabla\cdot\Big(D_mF(m^*,x)m^*(x)+\frac{\sigma^2}{2}\nabla m^*(x)\Big)=0$$

Theorem (Hu, R., Siska, Szpruch, '19)

Under mild conditions, if $m^* = \arg\min_m \left\{F(m) + \frac{\sigma^2}{2}\mathbf{Ent}(m)\right\}$ *then* m^* *is a stationary solution to the Fokker-Planck equation*

$$\partial_t m = \nabla\cdot\Big(D_mF(m,\cdot)m+\frac{\sigma^2}{2}\nabla m\Big) \tag{2}$$

Link to Overdamped Mean-field Langevin equation

The first order equation has a clear link to a Fokker-Planck equation.

$$\nabla\cdot\Big(D_mF(m^*,x)m^*(x)+\frac{\sigma^2}{2}\nabla m^*(x)\Big)=0$$

Theorem (Hu, R., Siska, Szpruch, '19)

Under mild conditions, if $m^*=\arg\min_m\left\{F(m)+\frac{\sigma^2}{2}\mathbf{Ent}(m)\right\}$ *then* m^* *is a stationary solution to the Fokker-Planck equation*

$$\partial_t m=\nabla\cdot\Big(D_mF(m,\cdot)m+\frac{\sigma^2}{2}\nabla m\Big) \tag{2}$$

It is well-known that the equation (2) characterizes the marginal law of the mean-field Langevin (MFL) dynamics:

$$dX_t=-D_mF(m_t,X_t)dt+\sigma dW_t,\qquad m_t=\mathrm{Law}(X_t)$$

Link to Underdamped Mean-field Langevin equation

Different from above, introduce the velocity variable V and consider the minimization:

$$\inf_{m=\mathrm{Law}(X,V)} F(m^X) + \frac{1}{2}\mathbb{E}^m\left[|V|^2\right] + \frac{\sigma^2}{2\gamma}\mathbf{Ent}(m)$$

Link to Underdamped Mean-field Langevin equation

Different from above, introduce the velocity variable V and consider the minimization:

$$\inf_{m=\mathrm{Law}(X,V)} F(m^X) + \frac{1}{2}\mathbb{E}^m\left[|V|^2\right] + \frac{\sigma^2}{2\gamma}\mathbf{Ent}(m)$$

The first order condition reads

$$D_m F(m^X, x) + \frac{\sigma^2}{2\gamma}\nabla_x \ln m(x,v) = 0 \quad \text{and} \quad v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m(x,v) = 0.$$

Link to Underdamped Mean-field Langevin equation

Different from above, introduce the velocity variable V and consider the minimization:

$$\inf_{m=\mathrm{Law}(X,V)} F(m^X) + \frac{1}{2}\mathbb{E}^m\left[|V|^2\right] + \frac{\sigma^2}{2\gamma}\mathbf{Ent}(m)$$

The first order condition reads

$$D_m F(m^X, x) + \frac{\sigma^2}{2\gamma}\nabla_x \ln m(x,v) = 0 \quad \text{and} \quad v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m(x,v) = 0.$$

One can again directly verify that the minimizer is an invariant measure of the underdamped MFL equation:

$$\begin{cases} dX_t = V_t, \\ dV_t = (-D_m F(m_t^X, X_t) - \gamma V_t)dt + \sigma dW_t \end{cases}$$

Difficulties: invariant measure of MFL

For the mean-field diffusion, the existence and uniqueness of the invariant measures is non-trivial, and the convergence of the marginal laws towards the invariant measure, if exists, is one of the long-standing problems in probability.

Difficulties: invariant measure of MFL

For the mean-field diffusion, the existence and uniqueness of the invariant measures is non-trivial, and the convergence of the marginal laws towards the invariant measure, if exists, is one of the long-standing problems in probability.

A simple example: mean-field Ornstein-Uhlenbeck process
$dX_t = (\alpha \mathbb{E}[X_t] - X_t)dt + dW_t$

- $\alpha < 1 \implies \exists$ unique invariant measure
- $\alpha > 1 \implies$ no invariant measure
- $\alpha = 1 \implies \exists$ multiple invariant measures

Difficulties: invariant measure of MFL

For the mean-field diffusion, the existence and uniqueness of the invariant measures is non-trivial, and the convergence of the marginal laws towards the invariant measure, if exists, is one of the long-standing problems in probability.

A simple example: mean-field Ornstein-Uhlenbeck process
$dX_t = (\alpha \mathbb{E}[X_t] - X_t)dt + dW_t$

- $\alpha < 1 \implies \exists$ unique invariant measure
- $\alpha > 1 \implies$ no invariant measure
- $\alpha = 1 \implies \exists$ multiple invariant measures

Given a convex F, the existence and uniqueness of the invariant measure of MFL is due to that of the minimizer m^*, thanks to the first order condition.

Difficulties: invariant measure of MFL

For the mean-field diffusion, the existence and uniqueness of the invariant measures is non-trivial, and the convergence of the marginal laws towards the invariant measure, if exists, is one of the long-standing problems in probability.

A simple example: mean-field Ornstein-Uhlenbeck process
$dX_t = (\alpha \mathbb{E}[X_t] - X_t)dt + dW_t$

- $\alpha < 1 \implies \exists$ unique invariant measure
- $\alpha > 1 \implies$ no invariant measure
- $\alpha = 1 \implies \exists$ multiple invariant measures

Given a convex F, the existence and uniqueness of the invariant measure of MFL is due to that of the minimizer m^*, thanks to the first order condition.

It remains to study the convergence of the marginal laws to the invariant measure.

Gradient flow and its analog

Define the energy functions for both overdamped and underdamped cases

$$U(m) = F(m) + \frac{\sigma^2}{2}\mathbf{Ent}(m), \quad \widehat{U}(m) = F(m^X) + \frac{1}{2}\mathbb{E}^m[|V|^2] + \frac{\sigma^2}{2\gamma}\mathbf{Ent}(m)$$

For the convergence towards the invariant measure, it is crucial to observe

Theorem (*Overdamped MFL*, Hu, R., Siska, Szpruch, '19)

$$dU(m_t) = -\mathbb{E}\left[\left|D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla_x \ln m_t(X_t)\right|^2\right] dt \quad \textit{for all } t > 0$$

Theorem (*Underdamped MFL*, Kazeykina, R., Tan, Yang, '20)

$$d\widehat{U}(m_t) = -\gamma\mathbb{E}\left[\left|V_t + \frac{\sigma^2}{2\gamma}\nabla_v \ln m_t(X_t, V_t)\right|^2\right] dt \quad \textit{for all } t > 0$$

Gradient flow and its analog

Define the energy functions for both overdamped and underdamped cases

$$U(m) = F(m) + \frac{\sigma^2}{2}\mathbf{Ent}(m), \quad \widehat{U}(m) = F(m^X) + \frac{1}{2}\mathbb{E}^m[|V|^2] + \frac{\sigma^2}{2\gamma}\mathbf{Ent}(m)$$

For the convergence towards the invariant measure, it is crucial to observe

Theorem (*Overdamped MFL*, Hu, R., Siska, Szpruch, '19)

$$dU(m_t) = -\mathbb{E}\left[\left|D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla_x \ln m_t(X_t)\right|^2\right] dt \quad \textit{for all } t > 0$$

Theorem (*Underdamped MFL*, Kazeykina, R., Tan, Yang, '20)

$$d\widehat{U}(m_t) = -\gamma\mathbb{E}\left[\left|V_t + \frac{\sigma^2}{2\gamma}\nabla_v \ln m_t(X_t, V_t)\right|^2\right] dt \quad \textit{for all } t > 0$$

Due to the generalized Itô calculus and time-reversal of diffusions.

Convergence for convex F

- *Overdamped:* $dU(m_t) = -\mathbb{E}\Big[\big|D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla_x \ln m_t(X_t)\big|^2\Big] dt$.
  Heuristically, $U(m_t)$ decreases till m_t hits m^* s.t.

$$D_m F(m^*, x) + \frac{\sigma^2}{2}\nabla_x \ln m^*(x) = 0$$

Convergence for convex F

- *Overdamped:* $dU(m_t) = -\mathbb{E}\Big[\Big|D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla_x \ln m_t(X_t)\Big|^2\Big]dt$.
  Heuristically, $U(m_t)$ decreases till m_t hits m^* s.t.

  $$m^* = \arg\min_m U(m)$$

  and $m_t \equiv m^*$ afterwards.

Convergence for convex F

- *Overdamped:* $dU(m_t) = -\mathbb{E}\Big[\big|D_m F(m_t, X_t) + \frac{\sigma^2}{2}\nabla_x \ln m_t(X_t)\big|^2\Big]dt.$
  Heuristically, $U(m_t)$ decreases till m_t hits m^* s.t.
  $$m^* = \arg\min_m U(m)$$
  and $m_t \equiv m^*$ afterwards.
- *Underdamped:* $d\widehat{U}(m_t) = -\gamma\mathbb{E}\Big[\big|V_t + \frac{\sigma^2}{2\gamma}\nabla_v \ln m_t(X_t, V_t)\big|^2\Big]dt$
  Similarly, $\widehat{U}(m_t)$ shall decrease till m_t hits m^* s.t.
  $$v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x, v) = 0$$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \ \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \ \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

We can prove that

- Overdamped: $m^* \in w(m_0) \Longrightarrow D_m F(m^*, x) + \frac{\sigma^2}{2}\nabla_x \ln m^*(x) = 0$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \ \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

We can prove that

- Overdamped: $m^* \in w(m_0) \Longrightarrow \ m^* = \arg\min_m U(m)$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \ \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

We can prove that

- Overdamped: $m^* \in w(m_0) \Longrightarrow \quad w(m_0) = \{\arg\min_m U(m)\}$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \ \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

We can prove that

- Overdamped: $m^* \in w(m_0) \Longrightarrow \quad m_t \to \arg\min_m U(m)$ in $\mathcal{W}_1$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

We can prove that

- Overdamped: $m^* \in w(m_0) \Longrightarrow \quad m_t \to \arg\min_m U(m)$ in $\mathcal{W}_1$
- Underdamped: $m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0$

A bit more rigorously...

The intuition above can be materialized by the LaSalle's invariance principle and the functional inequalities. Define the set of cluster points

$$w(m_0) := \{m : \ \exists (m_{t_n})_{n\in\mathbb{N}} \text{ s.t. } \lim_{n\to\infty} \mathcal{W}_1(m_{t_n}, m) = 0\}$$

Invariance Principle says:

$$\mathrm{Law}(X_0) \in w(m_0) \Longrightarrow \mathrm{Law}(X_t) \in w(m_0) \text{ for all } t > 0$$

We can prove that

- Overdamped: $m^* \in w(m_0) \Longrightarrow \quad m_t \to \arg\min_m U(m)$ in $\mathcal{W}_1$
- Underdamped: $m^* \in w(m_0) \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$.

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$dV_t h(X_t) = \Big(\big(\dot{h}(X_t)\cdot V_t\big)V_t + h(X_t)(-D_m F(m_t^X, X_t) - \gamma V_t)\Big)dt + dM_t$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$d\mathbb{E}\Big[V_t h(X_t)\Big] = \mathbb{E}\Big[\big(\dot h(X_t)\cdot V_t\big)V_t + h(X_t)(-D_m F(m_t^X, X_t) - \gamma V_t)\Big]dt$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$0 \quad = \quad \mathbb{E}\Big[\frac{\sigma^2}{2\gamma}\dot{h}(X_t) - h(X_t)D_m F(m_t^X, X_t)\Big]$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$\begin{aligned} 0 &= \mathbb{E}\Big[\frac{\sigma^2}{2\gamma}\dot{h}(X_t) - h(X_t)D_mF(m_t^X, X_t)\Big] \\ &= \mathbb{E}\Big[-\frac{\sigma^2}{2\gamma}h(X_t)\nabla_x \ln m_t(X_t, V_t) - h(X_t)D_mF(m_t^X, X_t)\Big] \end{aligned}$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$\begin{aligned} 0 &= \mathbb{E}\Big[\frac{\sigma^2}{2\gamma}\dot{h}(X_t) - h(X_t)D_mF(m_t^X, X_t)\Big] \\ &= \mathbb{E}\Big[h(X_t)\Big(-\frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t, V_t) - D_mF(m_t^X, X_t)\Big)\Big] \end{aligned}$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$\begin{aligned} 0 &= \mathbb{E}\Big[\frac{\sigma^2}{2\gamma}\dot{h}(X_t) - h(X_t)D_mF(m_t^X, X_t)\Big] \\ &= \mathbb{E}\Big[h(X_t)\Big(-\frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t, V_t) - D_mF(m_t^X, X_t)\Big)\Big] \\ \Rightarrow &\quad \frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t, V_t) + D_mF(m_t^X, X_t) = 0 \end{aligned}$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$\begin{aligned} 0 &= \mathbb{E}\Big[\frac{\sigma^2}{2\gamma}\dot{h}(X_t) - h(X_t)D_m F(m_t^X, X_t)\Big] \\ &= \mathbb{E}\Big[h(X_t)\Big(-\frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t,V_t) - D_m F(m_t^X, X_t)\Big)\Big] \\ &\Rightarrow \frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t,V_t) + D_m F(m_t^X, X_t) = 0 \\ &\Rightarrow m_t \equiv \arg\min_m \widehat{U}(m) \end{aligned}$$

Complete the proof for Underdamped MFL

Recall that

$$m^* \in w(m_0) \Longrightarrow v + \frac{\sigma^2}{2\gamma}\nabla_v \ln m^*(x,v) = 0 \Longrightarrow m^*(x,v) = g(x)e^{-\frac{\gamma}{\sigma^2}v^2}$$

Consider any smooth function h with compact support. Let $\mathrm{Law}(X_0) \in w(m_0)$. By Itô's formula,

$$\begin{aligned} 0 &= \mathbb{E}\Big[\frac{\sigma^2}{2\gamma}\dot{h}(X_t) - h(X_t)D_mF(m_t^X, X_t)\Big] \\ &= \mathbb{E}\Big[h(X_t)\Big(-\frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t, V_t) - D_mF(m_t^X, X_t)\Big)\Big] \\ \Rightarrow &\quad \frac{\sigma^2}{2\gamma}\nabla_x \ln m_t(X_t, V_t) + D_mF(m_t^X, X_t) = 0 \\ \Rightarrow &\quad w(m_0) = \{\arg\min_m \widehat{U}(m)\} \end{aligned}$$

Convergence rate for special case

For possibly non-convex F such that $D_m F(m, x)$ bearing small mean-field dependence, we can prove the contraction results using synchronous-reflection couplings.

Theorem (*Overdamped MFL*, Hu, R., Siska, Szpruch, '19)

$\mathcal{W}_1(m_t, m'_t) \leq Ce^{-\lambda t}\mathcal{W}_1(m_0, m'_0)$.

Theorem (*Underdamped MFL*, Kazeykina, R., Tan, Yang, '20)

$\mathcal{W}_\psi(m_t, m'_t) \leq Ce^{-\lambda t}\mathcal{W}_\psi(m_0, m'_0)$, *with the semi-metric*
$\mathcal{W}_\psi(m, m') = \inf\{\int \psi((x, v), (x', v'))\pi(dx, dy) : \ \pi \text{ is a coupling of } m, m'\}$

Convergence rate for special case

For possibly non-convex F such that $D_m F(m,x)$ bearing small mean-field dependence, we can prove the contraction results using synchronous-reflection couplings.

Theorem (*Overdamped MFL*, Hu, R., Siska, Szpruch, '19)

$\mathcal{W}_1(m_t, m'_t) \le Ce^{-\lambda t}\mathcal{W}_1(m_0, m'_0)$.

Theorem (*Underdamped MFL*, Kazeykina, R., Tan, Yang, '20)

$\mathcal{W}_\psi(m_t, m'_t) \le Ce^{-\lambda t}\mathcal{W}_\psi(m_0, m'_0)$, *with the semi-metric*
$\mathcal{W}_\psi(m, m') = \inf\{\int \psi((x,v),(x',v'))\pi(dx,dy) : \ \pi$ *is a coupling of* $m, m'\}$

Regretfully, the small mean-field dependence assumption is corresponding to the over-regularized problem in the context of neural networks.

Table of Contents

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known.

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known. Taking Wasserstein distance as example, we aim at sampling $\hat{\mu}$ by

$$\min_{\mu} \mathcal{W}_1(\mu, \hat{\mu})$$

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known. Taking Wasserstein distance as example, we aim at sampling $\hat{\mu}$ by

$$\begin{aligned} & \min_{\mu} \mathcal{W}_1(\mu, \hat{\mu}) \\ = \; & \min_{\mu} \sup_{f \in \mathrm{Lip}_1} \int f(x)(\mu - \hat{\mu})(dx) \end{aligned}$$

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known. Taking Wasserstein distance as example, we aim at sampling $\hat{\mu}$ by

$$\begin{aligned} & \min_{\mu} \mathcal{W}_1(\mu, \hat{\mu}) \\ = & \min_{\mu} \sup_{f \in \mathrm{Lip}_1} \int f(x)(\mu - \hat{\mu})(dx) \\ \approx & \min_{\mu} \sup_{f \in \mathcal{E}} \int f(x)(\mu - \hat{\mu})(dx) \end{aligned}$$

where $\mathcal{E} = \Big\{ z \mapsto \mathbb{E}^m\big[\varphi(X, z)\big] : \ X \sim m \in \mathcal{P}(\mathbb{R}^{n^2}) \Big\}$

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known. Taking Wasserstein distance as example, we aim at sampling $\hat{\mu}$ by

$$\min_{\mu} \sup_{f \in \mathcal{E}} \int f(x)(\mu - \hat{\mu})(dx)$$

where $\mathcal{E} = \left\{ z \mapsto \mathbb{E}^m\left[\varphi(X, z)\right] : \; X \sim m \in \mathcal{P}(\mathbb{R}^{n^2}) \right\}$

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known. Taking Wasserstein distance as example, we aim at sampling $\hat{\mu}$ by

$$\min_{\mu} \sup_{f \in \mathcal{E}} \int f(x)(\mu - \hat{\mu})(dx)$$

where $\mathcal{E} = \Big\{ z \mapsto \mathbb{E}^m \big[\varphi(X, z)\big] : \; X \sim m \in \mathcal{P}(\mathbb{R}^{n^2}) \Big\}$ GAN can be viewed as a zero-sum game between the generator and the discriminator:

$$\begin{cases} \text{Gen.}: & \inf_{\mu \in \mathcal{P}(\mathbb{R}^{n^1})} \int \mathbb{E}^m[\varphi(X, z)](\mu - \hat{\mu})(dz) + \frac{\sigma^2}{2}\big(\mathbf{Ent}(\mu) - \mathbf{Ent}(m)\big) \\ \text{Discr.}: & \inf_{m \in \mathcal{P}(\mathbb{R}^{n^2})} - \int \mathbb{E}^m[\varphi(X, z)](\mu - \hat{\mu})(dz) + \frac{\sigma^2}{2}\big(\mathbf{Ent}(m) - \mathbf{Ent}(\mu)\big) \end{cases}$$

GAN and zero-sum game

The Generative Adversary Network aims at sampling a target a probability measure $\hat{\mu} \in \mathcal{P}(\mathbb{R}^{n^1})$ only empirically known. Taking Wasserstein distance as example, we aim at sampling $\hat{\mu}$ by

$$\min_{\mu} \sup_{f \in \mathcal{E}} \int f(x)(\mu - \hat{\mu})(dx)$$

where $\mathcal{E} = \left\{ z \mapsto \mathbb{E}^m \left[\varphi(X, z) \right] : \; X \sim m \in \mathcal{P}(\mathbb{R}^{n^2}) \right\}$ GAN can be viewed as a zero-sum game between the generator and the discriminator:

$$\begin{cases} \text{Gen.}: & \inf_{\mu \in \mathcal{P}(\mathbb{R}^{n^1})} \int \mathbb{E}^m[\varphi(X,z)](\mu - \hat{\mu})(dz) + \frac{\sigma^2}{2}\left(\mathbf{Ent}(\mu) - \mathbf{Ent}(m)\right) \\ \text{Discr.}: & \inf_{m \in \mathcal{P}(\mathbb{R}^{n^2})} - \int \mathbb{E}^m[\varphi(X,z)](\mu - \hat{\mu})(dz) + \frac{\sigma^2}{2}\left(\mathbf{Ent}(m) - \mathbf{Ent}(\mu)\right) \end{cases}$$

In particular, $\mu, m \mapsto F(\mu, m) := \int \mathbb{E}^m[\varphi(X,z)](\mu - \hat{\mu})(dz)$ are linear.

The feedback of the generator

Due to the linearity, the solution to the generator given m (choice of discriminator) is explicit:

$$\mu^*[m](z) = C(m)^{-1} e^{-\frac{2}{\sigma^2}\left(\mathbb{E}^m[\varphi(X,z)]\right)},$$

where $C(m)$ is the normalization constant.

The feedback of the generator

Due to the linearity, the solution to the generator given m (choice of discriminator) is explicit:

$$\mu^*[m](z) = C(m)^{-1} e^{-\frac{2}{\sigma^2}\left(\mathbb{E}^m[\varphi(X,z)]\right)},$$

where $C(m)$ is the normalization constant. Therefore the value of the game can be rewritten as

$$\min_m \max_\mu -F(\mu, m) + \frac{\sigma^2}{2}\left(\mathbf{Ent}(m) - \mathbf{Ent}(\mu)\right) = \min_m G(m) + \frac{\sigma^2}{2}\mathbf{Ent}(m)$$

where $\quad G(m) := -F(\mu^*[m], m) - \dfrac{\sigma^2}{2}\mathbf{Ent}(\mu^*[m])$ is convex

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = -\int \mathbb{E}^m[\varphi(X, z)](\mu^* - \hat{\mu})(dz) - \frac{\sigma^2}{2}\mathbf{Ent}(\mu^*[m])$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = -\int \mathbb{E}^m[\varphi(X,z)](\mu^* - \hat{\mu})(dz) + \frac{\sigma^2}{2}\int \big(\ln C(m) + \frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]\big)\mu^*(dz)$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X, z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X, z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

Recall $C(m) = \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]}dz$, and thus

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X,z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

Recall $C(m) = \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]}dz$, and thus

$$\frac{\delta G}{\delta m} = \int \varphi(x,z)\hat{\mu}(dz) - \frac{\sigma^2}{2C(m)} \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]} \frac{2}{\sigma^2}\varphi(x,z)dz$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X, z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

Recall $C(m) = \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]}dz$, and thus

$$\frac{\delta G}{\delta m} = \int \varphi(x, z)\hat{\mu}(dz) - \frac{1}{C(m)} \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]}\varphi(x, z)dz$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X,z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

Recall $C(m) = \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]} dz$, and thus

$$\frac{\delta G}{\delta m} = \int \varphi(x,z)(\hat{\mu} - \mu^*[m])(dz)$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X, z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

Recall $C(m) = \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]}dz$, and thus

$$D_m G(m, x) \quad = \quad \int \nabla_x \varphi(x, z)(\hat{\mu} - \mu^*[m])(dz)$$

A convergent algorithm

Therefore the choice of the discriminator at the equilibrium is the invariant measure of the MFL dynamics:

$$dX_t = -D_m G(m_t, X_t)dt + \sigma dW_t$$

and the intrinsic derivative can be computed explicitly:

$$G(m) = \int \mathbb{E}^m[\varphi(X, z)]\hat{\mu}(dz) + \frac{\sigma^2}{2} \ln C(m)$$

Recall $C(m) = \int e^{-\frac{2}{\sigma^2}\mathbb{E}^m[\varphi(X,z)]}dz$, and thus

$$D_m G(m, x) = \int \nabla_x \varphi(x, z)(\hat{\mu} - \mu^*[m])(dz)$$

Finally note that $\mu^*[m]$ can be sampled by MCMC.

A toy example

Here we sample the law $\hat{\mu} = \exp(1)$ with $\mu_0 = \mathcal{N}(0, 1)$.

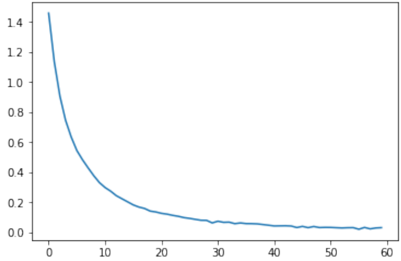

Figure: Pontential function value

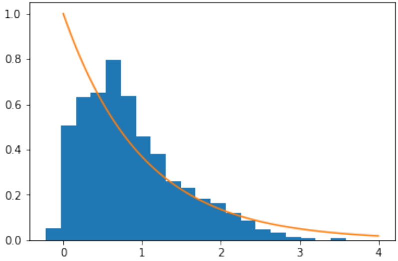

Figure: Histogram

Toy example with underdamped MFL

Similarly, we can train the discriminator by the underdamped MFL dynamics.

$$\begin{cases} dX_t = V_t \\ dV_t = \big(- D_m G(m_t^X, X_t) - \gamma V_t \big) dt + \sigma dW_t \end{cases}$$

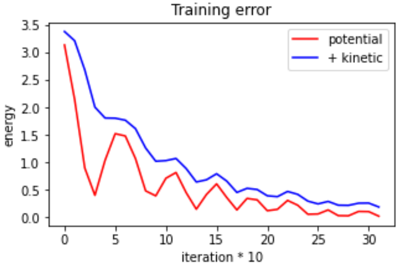

Figure: Energy value

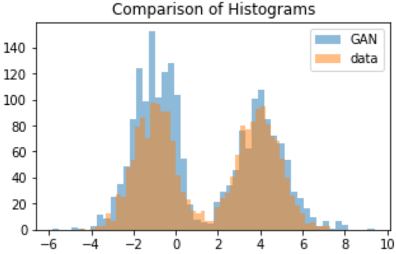

Figure: Histogram

Table of Contents

Optimization on random environment/with marginal constraint

More recently, with *G. Conforti* and *A. Kazeykina*, we discover that the previous analysis can be generalized to the optimization on random environment.

Optimization on random environment/with marginal constraint

More recently, with *G. Conforti* and *A. Kazeykina*, we discover that the previous analysis can be generalized to the optimization on random environment. Consider the optimization over $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{Y})$:

$$\min_{\pi:\ \pi_Y = \mathrm{m}} F(\pi) + \frac{\sigma^2}{2}\mathbf{Ent}(\pi | \mathrm{Leb} \times \mathrm{m})$$

where m is a fixed law on the environment $\mathbb{Y}$ (Polish).

First order condition

It is crucial to observe : for F convex we have

$$F(\pi') - F(\pi) \;\geq\; \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(\pi, x, \mathrm{y})(\pi' - \pi)(dx, dy)d\lambda$$

First order condition

It is crucial to observe : for F convex we have

$$F(\pi') - F(\pi) \;\geq\; \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(\pi, x, \mathrm{y})(\pi' - \pi)(dx, dy)d\lambda$$

Since $\pi_Y = \pi'_Y = \mathrm{m}$, a sufficient condition for m to be a minimizer is:

$\frac{\delta F}{\delta m}(\pi, x, \mathrm{y})$ does NOT depend on x

First order condition

It is crucial to observe : for F convex we have

$$F(\pi') - F(\pi) \;\geq\; \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(\pi, x, \mathrm{y})(\pi' - \pi)(dx, dy)d\lambda$$

Since $\pi_Y = \pi'_Y = \mathrm{m}$, a sufficient condition for m to be a minimizer is:

$$\nabla_x \tfrac{\delta F}{\delta m}(\pi, x, \mathrm{y}) = 0 \quad \text{for all } x, \text{ m-a.s. y}$$

First order condition

It is crucial to observe : for F convex we have

$$F(\pi') - F(\pi) \;\geq\; \int_{\mathbb{R}^d} \frac{\delta F}{\delta m}(\pi, x, \mathrm{y})(\pi' - \pi)(dx, dy)d\lambda$$

Since $\pi_Y = \pi'_Y = \mathrm{m}$, a sufficient condition for m to be a minimizer is:

$$\nabla_x \tfrac{\delta F}{\delta m}(\pi, x, \mathrm{y}) = 0 \quad \text{for all } x, \text{ m-a.s. y}$$

Theorem (Conforti, Kazeykina, R., '20)

Under mild conditions, if $\pi^* \in \arg\min_{\pi_Y = \mathrm{m}} \left\{ F(\pi) + \frac{\sigma^2}{2} \mathbf{Ent}(\pi | \mathrm{Leb} \times \mathrm{m}) \right\}$ *and let* $\pi^*(dx, dy) = \pi^*(x|\mathrm{y}) dx m(dy)$*, then*

$$\nabla_x \frac{\delta F}{\delta m}(\pi^*, x, \mathrm{y}) + \frac{\sigma^2}{2} \nabla_x \ln \pi^*(x|\mathrm{y}) = 0, \quad \textit{for all } x, \text{ m-}\textit{a.s.} \text{ y}. \tag{3}$$

Conversely, if F to be convex, (3) *implies m^* is the minimizer.*

Minimizer and invariant measure of MFL system

Due to the first order condition, the minimizer of V^σ is closely related to the invariant measure of the overdamped MFL system:

$$dX_t = -\nabla_x \frac{\delta F}{\delta m}(\pi_t, X_t, Y) + \sigma dW_t, \quad \pi_t = \mathrm{Law}(X_t, Y)$$

Minimizer and invariant measure of MFL system

Due to the first order condition, the minimizer of V^σ is closely related to the invariant measure of the overdamped MFL system:

$$dX_t^{\mathrm{y}} = -\nabla_x \frac{\delta F}{\delta m}(\pi_t, X_t^{\mathrm{y}}, \mathrm{y}) + \sigma dW_t, \quad \pi_t = \mathrm{Law}(X_t, Y)$$

Minimizer and invariant measure of MFL system

Due to the first order condition, the minimizer of V^σ is closely related to the invariant measure of the overdamped MFL system:

$$dX_t^{\mathrm{y}} = -\nabla_x \frac{\delta F}{\delta m}(\pi_t, X_t^{\mathrm{y}}, \mathrm{y}) + \sigma dW_t, \quad \pi_t = \mathrm{Law}(X_t, Y)$$

In particular, we know

- if F is convex, $\pi^* = \arg\min_{\pi_Y = \mathrm{m}} V^\sigma(\pi)$ iff π^* is the invariant measure
- for general F, if MFL system has unique invariant measure π^*, then $\pi^* = \arg\min_{\pi_Y = \mathrm{m}} V^\sigma(\pi)$

Minimizer and invariant measure of MFL system

Due to the first order condition, the minimizer of V^σ is closely related to the invariant measure of the overdamped MFL system:

$$dX_t^{\mathrm{y}} = -\nabla_x \frac{\delta F}{\delta m}(\pi_t, X_t^{\mathrm{y}}, \mathrm{y}) + \sigma dW_t, \quad \pi_t = \mathrm{Law}(X_t, Y)$$

In particular, we know

- if F is convex, $\pi^* = \arg\min_{\pi_Y = \mathrm{m}} V^\sigma(\pi)$ iff π^* is the invariant measure
- for general F, if MFL system has unique invariant measure π^*, then $\pi^* = \arg\min_{\pi_Y = \mathrm{m}} V^\sigma(\pi)$

Theorem (Conforti, Kazeykina, R., '20)

Under mild conditions, the MFL system admits unique strong solution and

$$dV^\sigma(\pi_t) = -\mathbb{E}\left[\left|\nabla_x \frac{\delta F}{\delta m}(\pi_t, X_t, Y) + \nabla_x \ln \pi_t(X_t|Y)\right|^2\right] dt, \quad \textit{for } t > 0$$

Convergence towards the invariant measure

- In case F convex, the V^σ again serves as Lyapunov function for the dynamic system (m_t). Provided that $\mathbb{Y}$ is countable or $\mathbb{R}^n$, we can show π_t converges to π^* in $\mathcal{W}_2$ based on Lasalle's invariant principle.

Convergence towards the invariant measure

- In case F convex, the V^σ again serves as Lyapunov function for the dynamic system (m_t). Provided that $\mathbb{Y}$ is countable or $\mathbb{R}^n$, we can show π_t converges to π^* in $\mathcal{W}_2$ based on Lasalle's invariant principle.
- For possibly non-convex but with small MF-dependence F, we can prove the contraction result:

Theorem (Conforti, Kazeykina, R., '20)

Under particular conditions, we have

$$\overline{\mathcal{W}}_1(\pi_t, \pi'_t) \le C e^{-\gamma t} \overline{\mathcal{W}}_1(\pi_0, \pi'_0),$$

$$\textit{where} \quad \overline{\mathcal{W}}_1(\pi, \pi') = \int \mathcal{W}_1\big(\pi(\cdot|\mathrm{y}), \pi'(\cdot|\mathrm{y})\big)\mathrm{m}(d\mathrm{y})$$

The constant γ can be computed, and once $\gamma > 0$ the MFL has a unique invariant measure, equal to the minimizer of V^σ, towards which the marginal laws converge.

Important example: Optimal control

Let $\mathbb{Y} = [0, T]$ and $\mathrm{m} = \mathrm{Leb}[0, T]$. Consider the relaxed optimal control

$$\inf_{\pi_Y = \mathrm{m}} \int_0^T \int L(\mathrm{y}, S_{\mathrm{y}}, x)\pi(x|\mathrm{y})d\mathrm{y} + g(S_T) + \frac{\sigma^2}{2}\mathbf{Ent}(\pi|\mathrm{Leb}),$$

$$\text{where} \quad S_{\mathrm{y}} = S_0 + \int_0^{\mathrm{y}} \int \varphi(u, S_u, x)\pi(x|u)du$$

Important example: Optimal control

Let $\mathbb{Y} = [0, T]$ and $\mathrm{m} = \mathrm{Leb}[0, T]$. Consider the relaxed optimal control

$$\inf_{\pi_Y = \mathrm{m}} \int_0^T \int L(\mathrm{y}, S_\mathrm{y}, x)\pi(x|\mathrm{y})d\mathrm{y} + g(S_T) + \frac{\sigma^2}{2}\mathbf{Ent}(\pi|\mathrm{Leb}),$$

$$\text{where} \quad S_\mathrm{y} = S_0 + \int_0^\mathrm{y} \int \varphi(u, S_u, x)\pi(x|u)du$$

Define the Hamiltonian function $H(\mathrm{y}, s, x, p) = L(\mathrm{y}, s, x) + p \cdot \varphi(\mathrm{y}, s, x)$.

Important example: Optimal control

Let $\mathbb{Y} = [0, T]$ and $\mathrm{m} = \mathrm{Leb}[0, T]$. Consider the relaxed optimal control

$$\inf_{\pi_Y = \mathrm{m}} \int_0^T \int L(\mathrm{y}, S_\mathrm{y}, x)\pi(x|\mathrm{y})d\mathrm{y} + g(S_T) + \frac{\sigma^2}{2}\mathbf{Ent}(\pi|\mathrm{Leb}),$$

$$\text{where} \quad S_\mathrm{y} = S_0 + \int_0^\mathrm{y} \int \varphi(u, S_u, x)\pi(x|u)du$$

Define the Hamiltonian function $H(\mathrm{y}, s, x, p) = L(\mathrm{y}, s, x) + p \cdot \varphi(\mathrm{y}, s, x)$.

We may compute $\frac{\delta F}{\delta m}(\pi, x, \mathrm{y}) = H(y, S_\mathrm{y}, x, P_\mathrm{y})$, where

$$P_\mathrm{y} = \nabla_s g(S_T) + \int_\mathrm{y}^T \int \nabla_s H(u, S_u, x, P_u)\pi(x|u)du$$

Important example: Optimal control

Let $\mathbb{Y} = [0, T]$ and $\mathrm{m} = \mathrm{Leb}[0, T]$. Consider the relaxed optimal control

$$\inf_{\pi_Y = \mathrm{m}} \int_0^T \int L(\mathrm{y}, S_\mathrm{y}, x)\pi(x|\mathrm{y})d\mathrm{y} + g(S_T) + \frac{\sigma^2}{2}\mathbf{Ent}(\pi|\mathrm{Leb}),$$

$$\text{where} \quad S_\mathrm{y} = S_0 + \int_0^\mathrm{y} \int \varphi(u, S_u, x)\pi(x|u)du$$

Define the Hamiltonian function $H(\mathrm{y}, s, x, p) = L(\mathrm{y}, s, x) + p \cdot \varphi(\mathrm{y}, s, x)$.

We may compute $\frac{\delta F}{\delta m}(\pi, x, \mathrm{y}) = H(y, S_\mathrm{y}, x, P_\mathrm{y})$, where

$$P_\mathrm{y} = \nabla_s g(S_T) + \int_\mathrm{y}^T \int \nabla_s H(u, S_u, x, P_u)\pi(x|u)du$$

The paper with *K. Hu and A. Kazeykina, '19* was devoted to this example and connect it to the deep neural network.

Deep neural network associated to relaxed controlled process

The Euler scheme introduces a forward propagation of a neural network:
$S_{t_{i+1}} \approx S_{t_i} + \frac{\delta t}{n_{t_{i+1}}} \sum_{j=1}^{n_{t_{i+1}}} \varphi(t_i, S_{t_i}, X^j_{t_{i+1}}, Z)$, where Z is the data.

Deep neural network associated to relaxed controlled process

The Euler scheme introduces a forward propagation of a neural network:
$S_{t_{i+1}} \approx S_{t_i} + \frac{\delta t}{n_{t_{i+1}}} \sum_{j=1}^{n_{t_{i+1}}} \varphi(t_i, S_{t_i}, X^j_{t_{i+1}}, Z)$, where Z is the data.

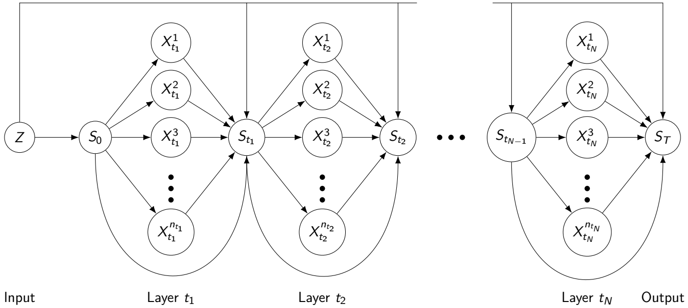

Figure: Neural network corresponding to the relaxed controlled process

Mean-field Langevin system ≈ Backward propagation

The architecture of the network is characterized by the average pooling after each layer!

Mean-field Langevin system ≈ Backward propagation

The architecture of the network is characterized by the average pooling after each layer!

The gradients of the parameters are easy to compute, due to the chain rule (or backward propagation):

$$X^{y}_{s_{j+1}} = X^{y}_{s_j} - \delta s \mathbf{E}\left[\nabla_a H(y, S^{y}_{s_j}, X^{y}_{s_j}, P^{y}_{s_j}, Z)\right] + \sigma \delta W_{s_j}, \text{ with } \delta s = s_{j+1} - s_j,$$

$$\text{where } P^{y_{i-1}}_s = P^{y_i}_s - \delta y \sum_{j=1}^{n_{y_{i+1}}} \nabla_s H\big(y_i, S^{y_i}_s, X^{y_i,j}_s, P^{y_i}_s, Z\big), \; P^T_s = \nabla_s g(S^T_s, Z),$$

(δW_{s_j}) are independent copies of $\mathcal{N}(0, \delta s)$.

Mean-field Langevin system ≈ Backward propagation

The architecture of the network is characterized by the average pooling after each layer!

The gradients of the parameters are easy to compute, due to the chain rule (or backward propagation):

$$dX_s^{\mathrm{y}} = -\mathbf{E}\big[\nabla_a H(\mathrm{y}, S_s^{\mathrm{y}}, X_s^{\mathrm{y}}, P_s^{\mathrm{y}}, Z)\big] ds + \sigma dW_s,$$

$$\text{where } dP_s^{\mathrm{y}} = -\mathbb{E}\Big[\nabla_s H\big(\mathrm{y}, S_s^{\mathrm{y}}, X_s^{\mathrm{y}}, P_s^{\mathrm{y}}, Z\big)\Big] d\mathrm{y},\ \ P_s^T = \nabla_s g(S_s^T, Z),$$

(δW_{s_j}) are independent copies of $\mathcal{N}(0, \delta s)$. Clearly, it is a discretization of the MF Langevin system.

Table of Contents

Nash equilibirum

Consider the game in which the i-th player chooses the probability measure π^i on $\mathbb{R}^{n^i} \times \mathbb{Y}$ as strategy to minimize his objective function $F^i(\pi^i, \pi^{-i})$, where π^{-i} is the joint strategy of other players on $\prod_{j \neq i} \mathbb{R}^{n^j} \times \mathbb{Y}$. We urge that the marginal law of π^i on $\mathbb{Y}$ is equal to the fixed law $m \in \mathcal{P}(\mathbb{Y})$.

Nash equilibirum

Consider the game in which the i-th player chooses the probability measure π^i on $\mathbb{R}^{n^i} \times \mathbb{Y}$ as strategy to minimize his objective function $F^i(\pi^i, \pi^{-i})$, where π^{-i} is the joint strategy of other players on $\prod_{j \neq i} \mathbb{R}^{n^j} \times \mathbb{Y}$. We urge that the marginal law of π^i on $\mathbb{Y}$ is equal to the fixed law $m \in \mathcal{P}(\mathbb{Y})$.

π^* is Nash eq: $\quad \pi^{*,i} \in \arg \min_{\pi^i:\ \pi^i_y = m} F^i(\pi^i, \pi^{*,-i}) + \frac{\sigma^2}{2} \mathbf{Ent}(\pi^i | \mathrm{Leb} \times m), \ \forall i$

Nash equilibirum

Consider the game in which the i-th player chooses the probability measure π^i on $\mathbb{R}^{n^i} \times \mathbb{Y}$ as strategy to minimize his objective function $F^i(\pi^i, \pi^{-i})$, where π^{-i} is the joint strategy of other players on $\prod_{j\neq i} \mathbb{R}^{n^j} \times \mathbb{Y}$. We urge that the marginal law of π^i on $\mathbb{Y}$ is equal to the fixed law $\mathrm{m} \in \mathcal{P}(\mathbb{Y})$.

$$\pi^* \text{ is Nash eq:} \quad \pi^{*,i} \in \arg \min_{\pi^i:\ \pi^i_y = \mathrm{m}} F^i(\pi^i, \pi^{*,-i}) + \frac{\sigma^2}{2} \mathbf{Ent}(\pi^i | \mathrm{Leb} \times \mathrm{m}), \ \forall i$$

Due to the previous first order condition we have

Theorem (Conforti, Kazeykina, R., '20)

If π is a Nash equilibrium, we have for $i = 1, \cdots, n$,

$$\nabla_{x^i} \frac{\delta F^i}{\delta \nu}(\pi^i, \pi^{-i}, x^i, \mathrm{y}) + \frac{\sigma^2}{2} \nabla_{x^i} \ln\left(\pi^i(x^i|\mathrm{y})\right) = 0 \quad \forall x^i \in \mathbb{R}^{n^i}, \ \mathrm{m}\text{-}a.s.\ \mathrm{y} \in \mathbb{Y}.$$

Uniqueness: Monotonicity condition

Theorem (Conforti, Kazeykina, R., '20)

Denote $\bar{x} = (x, y)$. The functions $(F^i)_{i=1,\cdots,n}$ satisfy the monotonicity condition, if for π, π' we have

$$\sum_{i=1}^{n} \int \left(\frac{\delta F^i}{\delta \nu}(\pi^i, \pi^{-i}, \bar{x}^i) - \frac{\delta F^i}{\delta \nu}(\pi'^i, \pi'^{-i}, \bar{x}^i) \right) (\pi - \pi')(d\bar{x}) \geq 0.$$

We have the following results:

(i) *if $n = 1$, a function F satisfies the monotonicity condition iff it is convex.*

(ii) *in general ($n \geq 1$), if $(F^i)_{i=1,\cdots,n}$ satisfy the monotonicity condition, then for any two Nash equilibria $\pi^*, \pi'^* \in \Pi$ we have $(\pi^*)^i = (\pi'^*)^i$ for all $i = 1, \cdots, n$.*

Proof of uniqueness

Sketch of proof: Since (F^i) is monotone,

$$\sum_{i=1}^{n} \int \left(\frac{\delta F^i}{\delta \nu}(\pi^i, \pi^{-i}, \bar{x}^i) - \frac{\delta F^i}{\delta \nu}(\pi'^i, \pi'^{-i}, \bar{x}^i) \right) (\pi - \pi')(d\bar{x}) \geq 0.$$

Proof of uniqueness

Sketch of proof: Since (F^i) is monotone,

$$\sum_{i=1}^{n} \int \left(\frac{\delta F^i}{\delta \nu}(\pi^i, \pi^{-i}, \bar{x}^i) - \frac{\delta F^i}{\delta \nu}(\pi'^i, \pi'^{-i}, \bar{x}^i) \right) (\pi - \pi')(d\bar{x}) \geq 0.$$

Together with the first order condition of equilibrium, we obtain

$$\sum_{i=1}^{n} \int \big(-\ln\big(\pi^i(x^i|y)\big) + \ln\big(\pi'^i(x^i|y)\big)\big)(\pi - \pi')(d\bar{x}) \geq 0.$$

Proof of uniqueness

Sketch of proof: Since (F^i) is monotone,

$$\sum_{i=1}^{n} \int \left(\frac{\delta F^i}{\delta \nu}(\pi^i, \pi^{-i}, \bar{x}^i) - \frac{\delta F^i}{\delta \nu}(\pi'^i, \pi'^{-i}, \bar{x}^i) \right) (\pi - \pi')(d\bar{x}) \geq 0.$$

Together with the first order condition of equilibrium, we obtain

$$-\sum_{i=1}^{n} \Big(\mathbf{Ent}(\pi^i|\pi'^i) + \mathbf{Ent}(\pi'^i|\pi^i) \Big) \geq 0.$$

Therefore $\pi^i = \pi'^i$ for all i.

Mean-field Langevin system and convergence to equilibrium

Again the FOC inspires the form of MFL dynamics:

$$dX_t^{i,\mathrm{y}} = \nabla_{x^i} \frac{\delta F^i}{\delta \nu}(\pi_t^i, \pi_t^{-i}, X_t^{i,\mathrm{y}}, \mathrm{y})dt + \sigma dW_t^i$$

Mean-field Langevin system and convergence to equilibrium

Again the FOC inspires the form of MFL dynamics:

$$dX_t^{i,\mathrm{y}} = \nabla_{x^i} \frac{\delta F^i}{\delta \nu}(\pi_t^i, \pi_t^{-i}, X_t^{i,\mathrm{y}}, \mathrm{y})dt + \sigma dW_t^i$$

In particular, if the game admits at least one Nash equilibrium and the MFL system has a unique invariant measure, then the invariant measure is an equilibrium.

Mean-field Langevin system and convergence to equilibrium

Again the FOC inspires the form of MFL dynamics:

$$dX_t^{i,\mathrm{y}} = \nabla_{x^i} \frac{\delta F^i}{\delta \nu}(\pi_t^i, \pi_t^{-i}, X_t^{i,\mathrm{y}}, \mathrm{y})dt + \sigma dW_t^i$$

In particular, if the game admits at least one Nash equilibrium and the MFL system has a unique invariant measure, then the invariant measure is an equilibrium.

- In the context of game, in general it is hard to find Lyapunov function.

Mean-field Langevin system and convergence to equilibrium

Again the FOC inspires the form of MFL dynamics:

$$dX_t^{i,y} = \nabla_{x^i} \frac{\delta F^i}{\delta \nu}(\pi_t^i, \pi_t^{-i}, X_t^{i,y}, y)dt + \sigma dW_t^i$$

In particular, if the game admits at least one Nash equilibrium and the MFL system has a unique invariant measure, then the invariant measure is an equilibrium.

- In the context of game, in general it is hard to find Lyapunov function.
- If the coefficient $\nabla_{x^i} \frac{\delta F^i}{\delta \nu}(\pi^i, \pi^{-i}, x^i, y)$ bears small mean-field dependence, we still can prove the contraction result, namely,

$$\overline{\mathcal{W}}_1(\pi_t, \pi_t') \le Ce^{-\gamma t}\overline{\mathcal{W}}_1(\pi_0, \pi_0')$$

Conclusion

References: One layer: Mei, Montanari, Nguyen '18, Hu, R., Siska, Szpruch '19; Deep/Neuron ODE: Hu, R., Kazeykina, '19, Jabir, Siska, Szpruch '19; Game on random environment: Conforti, R., Kazeykina, '20; Stochastic control: Siska, Szpruch, '20; Underdamped MFL: Kazeykina, R., Tan, Yang, '20 ...

Conclusion

References: One layer: Mei, Montanari, Nguyen '18, Hu, R., Siska, Szpruch '19; Deep/Neuron ODE: Hu, R., Kazeykina, '19, Jabir, Siska, Szpruch '19; Game on random environment: Conforti, R., Kazeykina, '20; Stochastic control: Siska, Szpruch, '20; Underdamped MFL: Kazeykina, R., Tan, Yang, '20 ...

- Mean-field Langevin dynamics is a natural model to analyze the (Hamiltonian) gradient descent for the overparametrized nonconvex optimization

Conclusion

References: One layer: Mei, Montanari, Nguyen '18, Hu, R., Siska, Szpruch '19; Deep/Neuron ODE: Hu, R., Kazeykina, '19, Jabir, Siska, Szpruch '19; Game on random environment: Conforti, R., Kazeykina, '20; Stochastic control: Siska, Szpruch, '20; Underdamped MFL: Kazeykina, R., Tan, Yang, '20 ...

- Mean-field Langevin dynamics is a natural model to analyze the (Hamiltonian) gradient descent for the overparametrized nonconvex optimization
- The calculus involving the measure derivatives characterizes the first order conditions, as well as allows the Itô-type calculus

Conclusion

References: One layer: Mei, Montanari, Nguyen '18, Hu, R., Siska, Szpruch '19; Deep/Neuron ODE: Hu, R., Kazeykina, '19, Jabir, Siska, Szpruch '19; Game on random environment: Conforti, R., Kazeykina, '20; Stochastic control: Siska, Szpruch, '20; Underdamped MFL: Kazeykina, R., Tan, Yang, '20 ...

- Mean-field Langevin dynamics is a natural model to analyze the (Hamiltonian) gradient descent for the overparametrized nonconvex optimization
- The calculus involving the measure derivatives characterizes the first order conditions, as well as allows the Itô-type calculus
- The relaxed control (continuous time or discrete time) can be viewed as an optimization with marginal constraint.

Thank you for your attention!