Mean-Field Optimisation Regularized by Fisher Information

Songbo WANG

CMAP, École Polytechnique

29/06/2022
BSDE 2022
Joint work with Julien CLAISSE, Giovanni CONFORTI, Zhenjie REN

Mean-field optimization

We consider a general "mean-field" function(al) $F : \mathcal{P}\left(\mathbb{R}^d\right) \to \mathbb{R}$. We study the optimization problem: $\inf_m F(m)$. Examples:

- Linear: $F(m) = \int f dm = \mathbb{E}_{X\sim m}\left[f(X)\right]$
- Quadratic: $F(m) = \int f dm + \int k(x,y) dm(x) dm(y)$
- Fancy: Neural networks

Neural networks

- One hidden layer
- $i = 1, \ldots, n$ – neurons
- $\varphi : \mathbb{R} \to \mathbb{R}$ – activation function, e.g. $\varphi(x) = x_+$ (ReLU)
- Quadratic cost

Problem: minimize

$$F_n(a, b, c) = \mathbf{E}\left[\left|f(Z) - \frac{1}{n}\sum_{i=1}^{n} c_k \varphi(a_k Z + b_k)\right|^2\right].$$

When $n \to \infty$,

$$F_n \to \mathbf{E}\left[\left|f(Z) - \mathbb{E}_m\left[C\varphi(AZ + B)\right]\right|^2\right] =: F(m)$$

where $(A, B, C) \sim m$.

Remarks: F is convex in m. It is no longer true when # layer ≥ 2.

Regularizations

Examples:

- entropy: $H(m) = H(m|e^{-U}) = \int (\log m + U) dm$
- Fisher information:
  $I(m) = \int \frac{|\nabla m|^2}{m} = \int |\nabla \log m|^2 dm = 4 \int |\nabla \sqrt{m}|^2 = 4 \|\nabla \sqrt{m}\|_{L^2(\mathbb{R}^d)}$

Regularized problem: $F^\sigma = F + \frac{\sigma^2}{2} H(m)$ or $F^\sigma = F + \frac{\sigma^2}{4} I(m)$.
Entropic case [Hu, Ren, Šiška, Szpruch, 2019]: the gradient descent w.r.t. $\mathcal{W}_2$ gives the marginal low of "mean-field Langevin"

$$dX_t = -DF(m_t, X_t)\, dt + \sigma dW_t, m_t \sim X_t.$$

m_t converges to the unique minimizer of $F^\sigma(m) = F(m) + \frac{\sigma^2}{2} H(m)$.
We consider the Fisher regularization in the following.

Mean-field C^1

Definition ("Functional", "flat", "L^2" derivative)

We say $F : \mathcal{P}\left(\mathbb{R}^d\right) \to \mathbb{R}$ is C^1 if there exists a continuous $\frac{\delta F}{\delta m} : \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$ s.t. for all $m_0, m_1 \in \mathcal{P}$

$$F(m_1) - F(m_0) = \int_0^1 \int \frac{\delta F}{\delta m}(m_t, x)\, d(m_1 - m_0)(x)\, dt$$

where $m_t = (1-t)\, m_0 + t m_1, t \in (0,1)$.

Remarks :

1. $\frac{\delta F}{\delta m}$ is defined up to a cst.
2. (F is convex). If m minimize F, then $\frac{\delta F}{\delta m}(m, \cdot)$ is cst.

First-order condition

Recall: $I(m) = \int \frac{|\nabla m|^2}{m}$.
We calculate formally:
$\delta I(m) = \int \frac{2\nabla m \cdot \nabla \delta m}{m} - \frac{|\nabla m|^2}{m^2}\delta m = \int \left(-2\nabla \cdot \left(\frac{\nabla m}{m}\right) - \frac{|\nabla m|^2}{m^2}\right)\delta m.$
Define

$$\frac{\delta F^\sigma}{\delta m} = \frac{\delta F}{\delta m} - \frac{\sigma^2}{2}\nabla \cdot \left(\frac{\nabla m}{m}\right) - \frac{\sigma^2}{4}\frac{|\nabla m|^2}{m^2}.$$

If F is convex, $F^\sigma = F + \frac{\sigma^2}{4} I$ is strictly convex and we expect

- if $\frac{\delta F^\sigma}{\delta m}(m_*, \cdot) = cst$, then m_* is the unique minimizer
- for all m_1, m_2, we have $F^\sigma(m_2) \geq F^\sigma(m_1) + \int \frac{\delta F^\sigma}{\delta m}(m_1, \cdot)(m_2 - m_1)$

Caveats:

- Fisher I is not strictly convex if the support of measures are disjoint
- $\frac{\delta F^\sigma}{\delta m}$ is singular and doesn't exist for general m s.t. $I(m) < +\infty$.

Observations

Denote $\psi = \sqrt{m}$. The FOC is equivalent to

$$cst = \frac{\delta F}{\delta m} - \sigma^2 \nabla \cdot \left(\frac{\nabla \psi}{\psi} \right) - \sigma^2 \frac{|\nabla \psi|^2}{\psi^2} = \frac{\delta F}{\delta m} - \sigma^2 \frac{\Delta \psi}{\psi}$$

$$\Leftrightarrow \quad cst \cdot \psi = \frac{\delta F}{\delta m} \psi - \sigma^2 \Delta \psi.$$

ψ is a eigenfunction of the mean-field Schrödinger operator

$$\sigma^2 \Delta - \frac{\delta F}{\delta m}(m, \cdot).$$

Denote $u = -\log m$. The FOC is equivalent to

$$cst = \frac{\delta F}{\delta m} + \frac{\sigma^2}{2} \Delta u - \frac{\sigma^2}{4} |\nabla u|^2.$$

It is a mean-field HJB equation associated to an ergodic control problem.

Definition of the dynamics

We consider the dynamics:

$$\partial_t m_t = -\frac{\delta F^\sigma}{\delta m}(m_t, \cdot)\, m_t$$

where $\frac{\delta F}{\delta m}$ is chosen such that $\int \frac{\delta F^\sigma}{\delta m}(m, x)\, dm = 0$.
Sanity check: $\partial_t \langle \mathbf{1}, m_t \rangle = 0$. Mass conserved.
Formally F^σ is decreasing:

$$\frac{dF^\sigma(m_t)}{dt} = -\int \left| \frac{\delta F^\sigma}{\delta m}(m_t, \cdot) \right|^2 dm_t$$

We can expect that $m_t \to$ the minimizer.

Equivalent Formulations

Dynamics

$$\frac{dm_t}{dt} = -\frac{\delta F^\sigma}{\delta m}(m_t, \cdot) m_t$$

Recall

$$\frac{\delta F^\sigma}{\delta m} = \frac{\delta F}{\delta m} - \frac{\sigma^2}{2} \nabla \cdot \left(\frac{\nabla m}{m} \right) - \frac{\sigma^2}{4} \frac{|\nabla m|^2}{m^2}.$$

The dynamics of $\psi = \sqrt{m}$: “mean-field dynamical Schrödinger”

$$\partial_t \psi_t = \frac{\sigma^2}{2} \Delta \psi_t - \frac{1}{2} \frac{\delta F}{\delta m} (m_t, \cdot) \, \psi_t$$

The dynamics of $u = -\log m$: “mean-field dynamical HJB”

$$\partial_t u = \frac{\sigma^2}{2} \Delta u - \frac{\sigma^2}{4} |\nabla u|^2 + \frac{\delta F}{\delta m} (m_t, \cdot)$$

Assumptions

F is continuous w.r.t. $\mathcal{W}_1$ and convex.
$F \in C^1$ and its derivative $\frac{\delta F}{\delta m}$ can decompose into

$$\frac{\delta F}{\delta m}(m,x) = g(x) + G(m,x)$$

where

1. $\kappa\,\mathrm{id} \leq \nabla^2 g \leq C\,\mathrm{id}$;
2. G is uniformly Lipschitz in x : $\sup_m \|\nabla G(m,\cdot)\|_\infty \leq L_G$.
3. ∇G is Lipschitz in m, x : $\forall m, m', x, x'$

$$\left|\nabla G(m,x) - \nabla G(m',x')\right| \leq L_G\left(\mathcal{W}_1(m,m') + |x-x'|\right).$$

Decomposition

$$\partial_t u = \frac{\sigma^2}{2}\Delta u - \frac{\sigma^2}{4}|\nabla u|^2 + \frac{\delta F}{\delta m}(m_t, \cdot)$$
$$= \frac{\sigma^2}{2}\Delta u - \frac{\sigma^2}{4}|\nabla u|^2 + g + G(m_t, \cdot)$$

We want to decompose the value function $u = v + w$ where v, w solves resp.

$$\partial_t v = \frac{\sigma^2}{2}\Delta v - \frac{\sigma^2}{4}|\nabla v|^2 + g$$
$$\partial_t w = \frac{\sigma^2}{2}\Delta w - \frac{\sigma^2}{2}\nabla v \cdot \nabla w - \frac{\sigma^2}{4}|\nabla w|^2 + G(m_t, \cdot)$$

Convexity of v

$$\partial_t v = \frac{\sigma^2}{2}\Delta v - \frac{\sigma^2}{4}|\nabla v|^2 + g.$$

The equation is classic (without mean-field). We have a classical solution. Moreover we have

Proposition

If $v_0 = v(0,\cdot)$ is θ_0-convex, then $v_t = v(t,\cdot)$ is θ_t-convex where θ_t solves Riccati:

$$\frac{d\theta_t}{dt} = \kappa - \frac{\sigma^2}{2}\theta_t^2$$

One proof: $dX_t = -\frac{\sigma^2}{2}\nabla v(T-t, X_t)dt + \sigma dW_t$, $Y_t = \nabla v(T-t, X_t)$, they solves FBSDE

$$dX_t = -\frac{\sigma^2}{2}Y_t dt + \sigma dW_t, X_0 = x$$
$$dY_t = -\nabla g(T-t, X_t)dt + Z_t dW_t, Y_T = \nabla v(0, X_T)$$

Consider two solutions $(X, Y), (X', Y')$, take the difference, use convexity...

A priori estimates of w

Recall that w solves

$$\partial_t w = \frac{\sigma^2}{2}\Delta w - \frac{\sigma^2}{2}\nabla v \cdot \nabla w - \frac{\sigma^2}{4}|\nabla w|^2 + G(m_t, \cdot)$$

Proposition

We suppose w solves classically on $[0, T]$

$$\partial_t w = \frac{\sigma^2}{2}\Delta w - \frac{\sigma^2}{2}\nabla v \cdot \nabla w - \frac{\sigma^2}{4}|\nabla w|^2 + L(t, x)$$

where L is uniformly Lipschitz in x and the initial value $w_0 = w(0, \cdot)$ is also Lipschitz. We suppose moreover $w, \nabla w$ is of polynomial growth. Then $\sup_{t \geq 0} \|\nabla w(t, \cdot)\|_\infty \leq C < +\infty$.

Ideas of proof

Write the optimal control problem

$$w(t,x) = \inf_{\alpha} \mathbb{E}\left[\int_0^t L(t-s, X_s) + \frac{\sigma^2}{4}|\alpha_s|^2 ds + w(0, X_t)\right]$$

$$dX_s = -\frac{\sigma^2}{2}(\alpha_s + \nabla v_{t-s}(X_s))\, ds + \sigma dW_s, \quad X_0 = x$$

Define X' starting from x', using the optimal control for x, and the same BM:

$$w(t,x') \leq \mathbb{E}\left[\int_0^t L(t-s, X'_s) + \frac{\sigma^2}{4}|\alpha_s|^2 ds + w(0, X'_t)\right]$$

$$dX'_s = -\frac{\sigma^2}{2}(\alpha_s + \nabla v_{t-s}(X'_s))\, ds + \sigma dW_s, \quad X'_0 = x'$$

X_t, X'_t becomes exponentially small thanks to the convexity of v. Then subtract...

Reflection coupling

Theorem (Eberle, 2011)

Let $b_1, b_2 : \mathbb{R}^d \to \mathbb{R}$, *of which* b_1 *is strictly decreasing:*

$$(x-y)\cdot(b_1(x)-b_1(y)) \leq -\theta\,|x-y|^2$$

and b_2 *is bounded.* $b = b_1 + b_2$. *If the diffusion* $dX_t = b(X_t)\,dt + dW_t$ *does not explode, then there exist csts* c, C *s.t. the marginals* m_t, m'_t *of the diffusion with* $m_0 = \delta_x, m'_0 = \delta_{x'}$ *satisfies*

$$\mathcal{W}_1\left(m_t, m'_t\right) \leq Ce^{-ct}\left|x - x'\right|.$$

Stability of ∇u

Proposition

Let $u_1 = v + w_1, u_2 = v + w_2$ be sums of form κ-convex + L-Lipschitz. Let $m_i = Z_i^{-1}\exp(-u_i)$. Then for a constant C depending only on κ, L, the bound holds

$$\mathcal{W}_1(m_1, m_2) \leq C \int |\nabla w_1 - \nabla w_2| dp_1$$

Ideas of proof: consider diffusion

$$dX_t = -\nabla w_i(X_t)dt + \sqrt{2}dW_t$$

and use Eberle's reflection coupling.

Stability

For a $f : \mathcal{R}^d \to \mathbb{R}, \alpha \geq 1$, define norm $\|f\|_{(\alpha)} := \sup_x \frac{|f(x)|}{(1+|x|^\alpha)}$.

Proposition

Suppose w_t, m_t $(\tilde{u}_t, \tilde{m}_t)$ solve

$$\partial_t w = \frac{\sigma^2}{2}\Delta w - \frac{\sigma^2}{2}\nabla v \cdot \nabla w - \frac{\sigma^2}{4}|\nabla w|^2 + G(m_t, \cdot) \textit{ resp. tilde version}$$

Then there exists a constant C_T such that

$$\|\nabla w_T - \nabla \tilde{w}_T\|_{(\alpha)} \leq C_T \left(\int_0^T \mathcal{W}_1(m_t, \tilde{m}_t) dt + \|\nabla w_0 - \nabla \tilde{w}_0\|_{(\alpha)} \right)$$

Estimate on second-order derivatives

Proposition

Let u solves for some flow of measures (m_t) on $\mathbb{R}_+$

$$\partial_t u = \frac{\sigma^2}{2}\Delta u - \frac{\sigma^2}{4}\left|\nabla u\right|^2 + \frac{\delta F}{\delta m}(m_t, \cdot)$$

then $\sup_{t\geq 0}\|\nabla^2 u_t\| < +\infty$.

Estimate on second-order derivatives: ideas of proof

∇u solves

$$\partial_t \nabla u = \frac{\sigma^2}{2} \Delta \nabla u - \frac{\sigma^2}{2} \nabla u \cdot \nabla^2 u + \nabla \frac{\delta F}{\delta m}$$

Probabilistic representation:

$$\nabla u(t,x) = \mathbb{E}\left[\int_0^t \nabla \frac{\delta F}{\delta m}(m_{t-s}, X_s) + \nabla u(0, X_t)\right]$$

$$\begin{aligned} dX_s &= -\sigma^2 \nabla u(t-s, X_s)\, ds + \sigma dW_s \\ &= -\sigma^2 (\nabla v + \nabla w)(t-s, X_s)\, ds + \sigma dW_s \end{aligned}$$

Drift = monotone + bounded. We use the reflection coupling to find a probability s.t. (X' follows the same diffusion whose starting point is x')

$$\mathbb{E}\left|X_s - X'_s\right| \leq Ce^{-cs}\left|x - x'\right|.$$

So ∇u is uniformly Lipschitz in x, i.e. $\sup_t \left\|\nabla^2 u_t\right\|_\infty < +\infty$.

Decrease of energy

Proposition

$$\frac{dF^\sigma(m_t)}{dt} = -\int \left| \frac{\delta F^\sigma}{\delta m}(m_t, x) \right|^2 m_t dx.$$

Tools: convexity, dominated convergence.
Convexity:

$$\int \frac{\delta F^\sigma}{\delta m}(m_{t+h}, x)(m_{t+h} - m_t)\, dx \geq F^\sigma(m_{t+h}) - F^\sigma(m_t)$$
$$\geq \int \frac{\delta F^\sigma}{\delta m}(m_t, x)(m_{t+h} - m_t)\, dx$$

where m_t solves classically the dynamics, i.e.

$$m_{t+h} - m_t = -\int_0^h \int \frac{\delta F^\sigma}{\delta m}(m_{t+r}, x)\, m_{t+r} dx dr.$$

Decrease of energy (continued)

To apply the dominated convergence, we need

1. $\sup_t \left|\frac{\delta F^\sigma}{\delta m}(m_t, x)\right| \le C\left(1 + |x|^2\right)$;
2. $\sup_t \int |x|^{4+\delta} m_t dx < +\infty$.

so that the integrand $\left|\frac{\delta F^\sigma}{\delta m}(m_t, x)\right|^2 m_t$ is bounded.
Recall: $\frac{\delta F^\sigma}{\delta m} = \frac{\delta F}{\delta m} + \frac{\sigma^2}{2}\Delta u - \frac{\sigma^2}{4}|\nabla u|^2$. Note that

1. We can prove ("turnpike" property, by Bernstein or BSDE) $\sup_t |\nabla v(x)| \le C(1 + |x|)$;
2. $\nabla u = \nabla v + \nabla w$ where ∇v is of linear growth, ∇w bounded;
3. $\sup_t \left\|\nabla^2 v_t\right\|_\infty < +\infty$;
4. $m_t = \exp(-v_t - w_t)$. We can use the concentration (or estimate directly the density) $\int |x|^p m_t dx < C_p$ for all $p \ge 1$.

Convergence

Theorem

$m_t \to m_*$ in L^1, where m_* is the unique minimizer to F^σ. Moreover, $\lim F^\sigma(m_t) = F^\sigma(m_*)$.

Ideas of proof:

- use structure of m_t (which follows from the estimates) to derive compactness
- use energy decrease formula and LaSalle's invariance principle to show all limit points $\hat{m}$ of m_t satisfy $\frac{\delta F^\sigma}{\delta m}(\hat{m}, \cdot) = 0$.
- for the convergence of energy,

$$F^\sigma(m_t) - F^\sigma(m_*) \le \int \frac{\delta F^\sigma}{\delta m}(m_t, \cdot)(m_t - m_*)$$

$$\le \left(\int \left| \frac{\delta F^\sigma}{\delta m}(m_t, \cdot) \right|^2 m_t \right)^{1/2} \left(\int \frac{(m_t - m_*)^2}{m_t} \right)^{1/2}$$

But caveats...

A gradient descent framework

Consider a C^1 convex $F: \mathbb{R}^d \to \mathbb{R}$, let $d(x,y) = \frac{1}{2}|x-y|^2$, $h > 0$. Define iteratively

$$y_{n+1} = \arg\min_y h^{-1} d(y, y_n) + F(y) \Leftrightarrow y_{n+1} = y_n - h\nabla F(y_{n+1})$$

In continuous time this becomes $\frac{dy}{dt} = -\nabla F(y)$, i.e. gradient descent. Generalizations to the space of measures:

- $F: \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and $d(m_1, m_2) = \mathcal{W}_2^2(m_1, m_2)$. This corresponds to the marginal of
$$\frac{dX_t}{dt} = -DF(X_t).$$
- $F^\sigma = F + \frac{\sigma^2}{2} H(m)$. $d = \mathcal{W}_2^2$. This corresponds to the marginal of
$$dX_t = -DF(X_t)dt + \sigma dW_t.$$
- $F^\sigma = F + \frac{\sigma^2}{2} H(m)$. $d(m_1, m_2) = H(m_1|m_2)$. [Liu, Majka, Szpruch, 2022]
- $F^\sigma = F + \frac{\sigma^2}{2} I(m)$. $d(m_1, m_2) = H(m_1|m_2)$.

Entropy-Fisher gradient descent

$d(m_1, m_2) = H(m_1|m_2)$, regularization by I.
At each step,

$$m^h_{k+1} = \arg\min_m h^{-1} H\left(m|m^h_k\right) + F^\sigma(m)$$

Formal first-order calculus:

$$\begin{aligned} 0 &= h^{-1}\delta \int \log \frac{m}{m^h_k} m + \delta F^\sigma(m) \\ &= h^{-1} \int \log \frac{m}{m^h_k} \delta m + \int \frac{\delta F^\sigma}{\delta m}(m, \cdot)\, \delta m \end{aligned}$$

so that

$$m^h_{k+1} = \frac{m^h_k}{Z_k} \exp\left(-h \frac{\delta F^\sigma}{\delta m}\left(m^h_{k+1}, \cdot\right)\right) \approx m^h_k \left(1 - h \frac{\delta F^\sigma}{\delta m}\left(m^h_{k+1}, \cdot\right)\right).$$

We expect $m^h_{kh} \to m_t$ when $h \to 0$ and $kh \to t$, where m_t solves

$$\frac{dm_t}{dt} = -\frac{\delta F^\sigma}{\delta m}(m_t, \cdot) m_t$$

Conclusions

1. Optimization problem with Fisher regularization (FOC)
2. Dynamics (MF Schrödinger, MF HJB, GD entropy-Fisher)
3. Convergence (no obvious rate – spectral inequalities destroyed by MF)
4. No numerics (for the moment)