Entropic Fictitious Play

for Mean-Field Optimization Problem

Songbo Wang

Joint work with Zhenjie Ren

École Polytechnique

2022-03-04

Outline

Outline

The Problem

- Minimize a known function $F : \mathcal{P}\left(\mathbb{R}^d\right) \to \mathbb{R}$
  - such functions are called "mean-field"...
- Examples:
  - Linear: $F(m) = \int f(x)\, m(dx)$
  - Quadratic: $F(m) = \frac{1}{2} \int \int k(x, y)\, m(dx)\, m(dy)$
  - Fancy: loss function of a neural network
- Entropic regularization: minimize $V^\sigma := F + \frac{\sigma^2}{2} H$
  - Fix a reference measure in Gibbs form: $R(dx) = \exp(-U(x))\, dx$
  - Entropy defined as $H(m) = H(m|R) = \int \log \frac{dm}{dR} m(dx)$
- Remarks: $H(P)$ is strictly convex, lower semi-continuous in P; gradient descent in $\mathcal{W}_2$ is a mean-field Langevin

Example: Single layer neural network

- Problem: minimize $\int \left(y - \frac{1}{n}\sum_{i=1}^{n} \beta_i \varphi\left(\alpha_i \cdot z + \gamma_i\right)\right)^2 \nu\left(dy, dz\right)$
- ν is an empirical measure, z feature, y label, φ activation, n number of neurons
- when $n \to \infty$, $\frac{1}{n}\sum_{i=1}^{n} \beta_i \varphi\left(\alpha_i \cdot z + \gamma_i\right) \to \mathbf{E}^m\left[\beta\varphi\left(\alpha \cdot z + \gamma\right)\right]$, $(\beta, \alpha, \gamma) \sim m$
- New problem: minimize $F(m) = \int \left(y - \mathbf{E}^m\left[\beta\varphi\left(\alpha \cdot z + \gamma\right)\right]\right)^2 \nu\left(dy, dz\right)$
- Lifting dimensions gives nice properties: F is convex
- However, if number of hidden layers $n \geq 2$, F is no longer convex

Outline

Calculus on the Space of Probabilities: a Primer

- Motivation: "differentiate" $F(m)$ against m
- Job: define $\frac{\delta F}{\delta m}$ such that
  $F((1-\varepsilon)m_0 + \varepsilon m_1) = F(m_0) + \varepsilon \left\langle \frac{\delta F}{\delta m}, m_1 - m_0 \right\rangle + o(\varepsilon)$
- Linear case: $F = \int f(x)\, m(dx) = \langle f, m \rangle$, $\frac{\delta F}{\delta m}$ should be f

Definition

A function $F : \mathcal{P}(\mathbb{R}^d) \to \mathbb{R}$ is called C^1 if there exists a bounded continuous function $\frac{\delta F}{\delta m} : \mathcal{P}(\mathbb{R}^d) \times \mathbb{R}^d \to \mathbb{R}$ such that

$$F(m_1) - F(m_0) = \int \int_0^1 \frac{\delta F}{\delta m}((1-t)m_0 + tm_1, x)\,(m_1(dx) - m_0(dx))$$

for all $m_0, m_1 \in \mathcal{P}(\mathbb{R}^d)$. The function $\frac{\delta F}{\delta m}$ is called functional derivative.

Calculus on the Space of Probabilities: Remarks

- The functional derivative is defined up to a constant (which may depend on m)
  - if $F \in C^1$ has functional derivative $\frac{\delta F}{\delta m}$
  - then $\frac{\delta F}{\delta m}(m, x) + \mathrm{const}(m)$ is also a functional derivative
- Quadratic example:
  - $F(m) = \frac{1}{2} \int \int k(x, y)\, m(dx)\, m(dy)$ with k bounded continuous
  - Then $\frac{\delta F}{\delta m}(m, x) = \int k(x, y)\, m(dy)$ is a possible function derivative
  - But any $\int k(x, y)\, m(dy) + G(m)$ with bounded continuous $G : \mathcal{P}\left(\mathbb{R}^d\right) \to \mathbb{R}$ is also a functional derivative
- For the function F in interest, we always fix ONE functional derivative $\frac{\delta F}{\delta m}$

Outline

First Order Necessary Condition

- Let $m^\star$ minimizes $V^\sigma = F + \frac{\sigma^2}{2}H$, what to say about $m^\star$?
- Natural candidate: $\frac{\delta V^\sigma}{\delta m}(m^\star, x) = \text{const}$
  - const instead of 0 is due to the ambiguity of functional derivative
- Problem: F is usually C^1, but H is for most cases not
- Formal calculations:
  - $H(m) = \int \log \frac{dm}{dR} m(dx) = \int m(x)(\log m(x) + U(x))\,dx$
  - $\delta H(m) = \delta \int m(x)(\log m(x) + U(x))\,dx =$ $\int \delta(m(x)(\log m(x) + U(x)))\,dx =$ $\int (\log m(x) + 1 + U(x))\,\delta m(x)\,dx$
  - Note $\int 1 \delta m(x)\,dx = 0$ (ambiguity of functional derivative strikes again)
  - Candidate (?): $\frac{\delta H}{\delta m}(m, x) = \log m(x) + U(x)$
  - Does not fit in the definition: $m(x)$ may be unbounded and discontinuous

First Order Necessary Condition: Assumptions

Let $p \geq 1$.

Assumption

$F \in C^1$ and is bounded from below.

Assumption

$R = \exp\left(-U\left(x\right)\right) dx$ is such that $\operatorname{ess\,inf}_{x \in \mathbb{R}^d} U\left(x\right) > -\infty$ and $\operatorname{ess\,lim\,inf}_{x \to \infty} \frac{U(x)}{|x|^p} > 0$.

Definition

$\mathcal{P}_p\left(\mathbb{R}^d\right) := \left\{m \in \mathcal{P}\left(\mathbb{R}^d\right) : \int |x|^p \, m\left(dx\right) < +\infty\right\}.$

First Order Necessary Condition: Result

Proposition

If $m^\star \in \mathcal{P}\left(\mathbb{R}^d\right)$ *minimizes* $V^\sigma = F + \frac{\sigma^2}{2} H$*, then* $m^\star \in \mathcal{P}_p\left(\mathbb{R}^d\right)$ *and has density w.r.t. Lebesgue. Moreover, the density satisfies*

$$\frac{\delta F}{\delta m}(m^\star, x) + \frac{\sigma^2}{2}\left(\log m^\star(x) + U(x)\right) = \mathit{const}, \quad \text{Lebesgue a.e.}$$

This validates our formal calculations!

Outline

Entropic Fictitious Play: Motivations

- We look for m such that $\frac{\delta F}{\delta m}(m,x) + \frac{\sigma^2}{2}(\log m(x) + U(x)) = \text{const}$
- First order condition (FOC) viewed as fixed point problem:
  - Define $\hat{m}$ by $\frac{\delta F}{\delta m}(m,x) + \frac{\sigma^2}{2}(\log \hat{m}(x) + U(x)) = \text{const}$
  - In Gibbs form: $\hat{m}(x) = \frac{1}{Z}\exp\left(-U(x) - \frac{2}{\sigma^2}\frac{\delta F}{\delta m}(m,x)\right)$
  - the mapping $m \mapsto \hat{m}$ has fixed point $m^\star$ iff $m^\star$ satisfies FOC
  - resembles Nash equilibrium
- Motivated, we consider the dynamics:

$$\frac{dm_t}{dt} = \alpha(\hat{m}_t - m_t)$$

  - α is a positive constant
  - resembles fictitious play in game theory

Wasserstein Distance

Let $p \geq 1$.

Definition

(M, d) metric space. p-Wasserstein is a distance between Borel probabilities in $\mathcal{P}_p(M)$ such that

$$\mathcal{W}_p(P, Q) = \inf_{X \sim P, Y \sim Q} \mathbf{E}\left[d(X, Y)^p\right]^{\frac{1}{p}}.$$

The inf is taken over all possible "couplings" of P and Q.

Fact

$\mathcal{W}_p$ metrizes the weak topology with convergent p-moment of $\mathcal{P}_p$, and $(\mathcal{P}_p, \mathcal{W}_p)$ is complete.

Entropic Fictitious Play: Wellposedness

From now on we fix a $1 \leq p \leq 2$.

Assumption

$\frac{\delta F}{\delta m}(m, x)$ is jointly Lipschitz in m, x, where the difference of m is measured by the p-Wasserstein distance $\mathcal{W}_p$.

Proposition

The dynamics

$$\frac{dm_t}{dt} = \alpha\left(\hat{m}_t - m_t\right) \tag{1}$$

is wellposed in $\mathcal{P}_p\left(\mathbb{R}^d\right)$, i.e. there exists a unique dynamics in $C\left([0, +\infty); (\mathcal{P}_p, \mathcal{W}_p)\right)$ solving eq. (1) for any initial value $m_0 \in \mathcal{P}_p$. Moreover we have continuous dependency on m_0.

Entropic Fictitious Play: Further Regularities

Proposition

If in addition to $m_0 \in \mathcal{P}_p$, the initial value m_0 has density w.r.t. Lebesgue, then the dynamics $(m_t)_t$ has also density for all $t > 0$. Moreover the function $t \mapsto m_t$ is C^1 and satisfies

$$\frac{dm_t(x)}{dt} = \alpha\left(\hat{m}_t(x) - m_t(x)\right)$$

for all x and all $t > 0$.

Entropic Fictitious Play: Convergence

- We would like to show m_t converges to some $m^\star$ satisfying the first order condition
- Formal calculations show that V^σ serves as a Lyapunov function

$$\frac{dV^\sigma(m_t)}{dt} = -\frac{\alpha\sigma^2}{2}\left(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\right)$$

- At least formally, V^σ decreases along $(m_t)_t$
- Since V^σ is finite, we hope $\lim_{t\to\infty} \frac{dV^\sigma(m_t)}{dt} = \lim_{t\to\infty} -\frac{\alpha\sigma^2}{2}\left(H(m_t|\hat{m}_t) + H(\hat{m}_t|m_t)\right) = 0$
- If we suppose $m_t \to$ some $m^\star$, by continuity of $\cdot \mapsto \hat{\cdot}$, $\hat{m}_t \to \widehat{m^\star}$
- Using again the semi-continuity of $H(\cdot|\cdot)$, we wish to have $H\left(m^\star|\widehat{m^\star}\right) = H\left(\widehat{m^\star}|m^\star\right) = 0$, i.e. $m^\star = \hat{m}^\star$, FOC satisfied

Entropic Fictitious Play: Convergence

If we suppose additionally

Assumption

The mapping $\cdot \mapsto \hat{\cdot}$ admits unique fixed point $m^\star$.

Assumption

The initial condition $m_0 \in \mathcal{P}_{p'}\left(\mathbb{R}^d\right)$ for some $p' > p$ and have finite entropy $H\left(m_0\right) < +\infty$

then we have

Theorem (Convergence)

$\lim_{t\to\infty} \mathcal{W}_p\left(m_t, m^\star\right) = 0$, and $\lim_{t\to\infty} m_t\left(x\right) = m^\star\left(x\right)$ for x a.e. The time derivative satisfies

$$\frac{dV^\sigma\left(m_t\right)}{dt} = -\frac{\alpha\sigma^2}{2}\left(H\left(m_t|\hat{m}_t\right) + H\left(\hat{m}_t|m_t\right)\right),$$

and its value also converges: $\lim_{t\to\infty} V^\sigma\left(m_t\right) = V^\sigma\left(m^\star\right)$.

Entropic Fictitious Play: Convex case

Assumption

F is convex and C^2, i.e. $\frac{\delta F}{\delta m} \in C^1$.

- $V^\sigma = F + \frac{\sigma^2}{2} H$ is strictly convex, since H is strictly convex
- Uniqueness of fixed point $m^\star$ of $\cdot \mapsto \hat{\cdot}$ follows automatically
- Rate of convergence can also be obtained:

Theorem

$$0 \leq V^\sigma(m_t) - V^\sigma(m^\star) \leq \frac{\sigma^2}{2} H(m_0|\hat{m}_0)\, e^{-\alpha t}.$$

Outline

Numerical Test: Samping $\hat{m}$

- $\hat{m}$ defined in Gibbs form: $\hat{m}(x) = \frac{1}{Z} \exp\left(-U(x) - \frac{2}{\sigma^2} \frac{\delta F}{\delta m}(m, x)\right)$
- To sample it, we note that it is the unique invariant measure of the Langevin dynamics

$$dX_t = -\left(\nabla \frac{\delta F}{\delta m}(m, x) + \frac{\sigma^2}{2} U(x)\right) dt + \sigma dB_t$$

  - under conditions on U, F...
- Langevin dynamics allows us to compute $\hat{m}$

Numerical Test: Result

We learn a 1d function $y = \cos 2\pi z, z \in [0, 1]$