# Optimisation champ-moyen régularisé par l'information de Fisher

Songbo WANG

CMAP, École Polytechnique

31/08/2022
MAS 2022
Joint work with Julien CLAISSE, Giovanni CONFORTI, Zhenjie REN

1 Motivations

2 Dynamique

3 Descente de gradient

# Optimisation champ-moyen

On étudie le problème d'optimisation : $\inf_m F(m)$ où $F : \mathcal{P}\left(\mathbb{R}^d\right) \to \mathbb{R}$.

Exemples :

- linéaire: $F(m) = \int f dm = \mathbb{E}_{X\sim m}[f(X)]$
- quadratique: $F(m) = \int f dm + \int k(x,y) dm(x) dm(y)$
- réseaux de neurons (NN)

L'exemple le plus simple de NN: une couche cachée de $n$ neurones.

But : minimiser

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

où $\varphi : \mathbb{R} \to \mathbb{R}$ est la fonction d'activition non-linéaire.

Quand $n \to \infty$,

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

où $(A,B,C) \sim m$.

Remarque : $F$ est convexe $m$. Ce n'est plus vrai pour des NN profonds...

# Régularisation

Pour $m \in \mathcal{P}(\mathbb{R}^d)$, quelques choix de régularisateurs :

- entropie : $H(m) = H(m|e^{-U}) = \int (\log m + U) dm$, où $U : \mathbb{R}^d \to \mathbb{R}$
- information de Fisher (p.r.à Leb):
  $I(m) = \int \frac{|\nabla m|^2}{m} = \int |\nabla \log m|^2 dm = 4 \int |\nabla \sqrt{m}|^2 = 4 \|\nabla \sqrt{m}\|_{L^2(\mathbb{R}^d)}$

Problème régularisé : $F^\sigma = F + \frac{\sigma^2}{2} H(m)$ or $F^\sigma = F + \frac{\sigma^2}{4} I(m)$.
Cas entropique [Hu, Ren, Šiška, Szpruch, 2019]: la descente de gradient p.r.à. $\mathcal{W}_2$ donne la loi marginale de Langevin champ-moyen

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

$m_t$ converge vers l'unique minimiseur de $F^\sigma(m) = F(m) + \frac{\sigma^2}{2} H(m)$.
On considère la régularisation de Fisher dans la suite.

# Dérivation d'une fonction champ-moyen

## Définition (Dérivée "functionelle", "plate", "$L^2$")

On dit $F : \mathcal{P}\left(\mathbb{R}^d\right) \to \mathbb{R}$ est $C^1$ s'il existe une continue $\frac{\delta F}{\delta m} : \mathcal{P}\left(\mathbb{R}^d\right) \times \mathbb{R}^d \to \mathbb{R}$ t.q. pour tout $m_0, m_1 \in \mathcal{P}$

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

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

Remarques :

1. $\frac{\delta F}{\delta m}$ est définie à cste près.
2. Si $F$ est convexe et si $m$ minimise $F$, alors $\frac{\delta F}{\delta m}(m, \cdot)$ est cste.

# Condition du premier ordre

Rappel : $I(m) = \int \frac{|\nabla m|^2}{m}$.
On calcule formellement :
$\delta I(m) = \int \frac{2\nabla m \cdot \nabla \delta m}{m} - \frac{|\nabla m|^2}{m^2}\delta m = \int \left(-2\nabla \cdot \left(\frac{\nabla m}{m}\right) - \frac{|\nabla m|^2}{m^2}\right)\delta m.$
Définissions

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

Si $F$ est convexe, $F^\sigma = F + \frac{\sigma^2}{4} I$ est aussi convexe et on espère

- si $\frac{\delta F^\sigma}{\delta m}(m_*, \cdot) = cste$, alors $m_*$ est l'unique minimiseur
- pour tout $m_1, m_2$, on a $F^\sigma(m_2) \geq F^\sigma(m_1) + \int \frac{\delta F^\sigma}{\delta m}(m_1, \cdot)(m_2 - m_1)$

Sauf que ...

- Fisher $I$ n'est pas strictement convexe si le support de deux mesures sont disjoint
- $\frac{\delta F^\sigma}{\delta m}$ est singulier et n'existe pas pour $m$ quelconque t.q. $I(m) < +\infty$.

1 Motivations

2 Dynamique

3 Descente de gradient

# Observations

Notons $\psi = \sqrt{m}$. La condition du premier ordre est équivalente à

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

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

$\psi$ est une fonction propre de l'opérateur de Schrödinger

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

Notons $u = -\log m$. La condition du premier ordre est équivalente à

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

C'est une équation HJB champ-moyen associée à un problème de contrôle ergodique.

# Définition de la dynamique

On considère la dynamique :

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

où $\frac{\delta F}{\delta m}$ est choisi t.q. $\int \frac{\delta F^\sigma}{\delta m}(m, x)\, dm = 0$.
Contrôle "sanitaire" : $\partial_t \langle \mathbf{1}, m_t \rangle = 0$. La masse est conservée.
Au niveau formel $F^\sigma$ décroit :

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

On espère alors que $m_t \to$ le minimiseur.

# Formulations équivalentes

Dynamique

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

Rappelons

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

La dynamique de $\psi = \sqrt{m}$ : Schrödinger dynamique champ-moyen

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

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

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

# Hypothèses

$F$ est continue p.r.à. $\mathcal{W}_1$ et convexe.
$F \in C^1$ en mesure et sa dérivée $\frac{\delta F}{\delta m}$ peut se décomposer

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

où

1. $\kappa\,\mathrm{id} \leq \nabla^2 g \leq C\,\mathrm{id}$;
2. $G$ est uniformément lipschitzienne en x : $\sup_m \|\nabla G(m,\cdot)\|_\infty \leq L_G$.
3. $\nabla G$ est lipschitzienne en $m, x$ : $\forall m, m', x, x'$

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

La valeur initiale $u_0$ se décompose $u_0 = v_0 + w_0$ où

1. $0 < \theta_0 \leq \nabla^2 v_0(x) \leq C < +\infty$
2. $w_0$ est lipschitzienne

# Décomposition de $u$

On décompose la solution de la HJB $u = v + w$ où $v, w$ résolvent résp.

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

Comparons à l'équation de $u$:

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

L'évolution de $v$ ne dépend pas de $m_t$. On se concentre à l'étude de $w$.

# Estimée a priori

## Théorème

*Supposons les hypothèses mentionnées sur $F$ et $u_0$. Soit $u : [0, T) \times \mathbb{R}^d \to \mathbb{R}$ une solution classique à la HJB champ-moyen de croissance polynomiale. Il existe alors $v_t, w_t$ résolvant les équations correspondantes t.q. $u_t = v_t + w_t$, et*

$$0 < \theta \leq \sup_{t,x} |\nabla^2 v(t,x)| < +\infty, \sup_{t,x} |\nabla w(t,x)| \leq L < +\infty$$

*où $\theta, L$ sont des cstes indépendant de $T$. De plus, les dérivées secondes spatiales de $u$ peut aussi être borné uniformément en temps :*

$$\sup_{t,x} |\nabla^2 u(t,x)| \leq C, \text{ where } C \text{ is independent of } T.$$

Idées de la preuve : utiliser contrôle optimal et construire des couplages. Pour borner $\nabla^2 u$ on utilise la couplage par réflexion d'Eberle.

# Le caractère bien posé

## Théorème (Existence)

*Il existe des solutions classiques $v_t, w_t$ aux équations correspondantes.*

Soit $\alpha > 1$. Considérons la norme $\alpha$-croissance de fonctions à valeur vectorielle $f : \mathbb{R}^d \to X$: $\|f\|_\alpha = \sup_x \frac{|f(x)|}{1+|x|^\alpha}$.

## Théorème (Stabilité)

*Si $u_t, u_t^i$ résolvent la HJB dynamique classiquement et $\lim_i \|\nabla u_0^i - \nabla u_0\|_\alpha = 0$, alors pour tout $t$, $\lim_i \|\nabla u_t^i - \nabla u_t\|_\alpha = 0$.*

Idées de la preuve : construire l'application

$$(w_t)_{t\in[0,T]} \mapsto (m_t)_{t\in[0,T]} \mapsto (w'_t)_{t\in[0,T]}$$

et appliquer un argument de point fixe de Banach. On obtient une $L^1$ inégalité qui borne la cste lip du premier composant de l'application $w \mapsto m$.

# $L^1$ inégalité

## Proposition

*Soit* $\nu \in \mathcal{P}(\mathbb{R}^d)$ *t.q.*

$$-\log \nu(x) = v(x) + w(x) \in C^2$$

*où* $v$ *est* $\theta$*-convexe et* $w$ *est* $L$*-lip,* $\kappa, L > 0$*. Alors il existe des cstes* $C(\theta, L)$ *t.q. pour tout* $\mu \in \mathcal{P}_1(\mathbb{R}^d)$ *on a*

$$\mathcal{W}_1(\mu, \nu) \leq C(\theta, L) \int \left| \log \frac{d\mu}{d\nu} \right| d\mu$$

Remarque : la version $L^2$ de cette inégalité est Talagrand + log-Sobolev. Mais pour que log-S soit vérifiée $w$ devrait être bornée (Bakry-Emery) au lieu de lip.

Idées de la preuve : considérer les processus de diffusion couplés par réflexion

$$dX_t = \nabla \log \nu(X_t)dt + \sqrt{2}dW_t, dX'_t = \nabla \log \mu(X'_t)dt + \sqrt{2}dW'_t$$

# Convergence

## Proposition

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

Idées de la preuve : utiliser les estimées sur $u_t$ pour vérifier que l'intégrale est bien défini et appliquer le théorème de convergence dominée.

## Théorème

*$m_t \to m_*$ en $L^1$, où $m_*$ est l'unique minimiseur de $F^{\sigma}$. De plus,* $\lim F^{\sigma}(m_t) = F^{\sigma}(m_*)$.

Idées de la preuve : compacité en $L^1$ (grâce aux estimées sur $u$) et le principe de LaSalle.

1 Motivations

2 Dynamique

3 Descente de gradient

# Un cadre de descente de gradient

Considérons une $C^1$ convexe $F : \mathbb{R}^d \to \mathbb{R}$. Soit $d(x,y) = \frac{1}{2}|x-y|^2$, $h > 0$. Définissons par itération

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

En temps continu cela devient $\frac{dy}{dt} = -\nabla F(y)$, i.e. la descente de gradient. Généralisations dans l'espace de mesures :

- $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ and $d(m_1, m_2) = \mathcal{W}_2^2(m_1, m_2)$. Cela correspond à la loi marginale de
$$\frac{dX_t}{dt} = -DF(X_t).$$
- $F^\sigma = F + \frac{\sigma^2}{2} H(m)$. $d = \mathcal{W}_2^2$. [Jordan, Kinderlehrer, Otto, 1998] Cela correspond à la loi marginale de
$$dX_t = -DF(X_t)dt + \sigma dW_t.$$
- $F^\sigma = F + \frac{\sigma^2}{2} H(m)$. $d(m_1, m_2) = H(m_1|m_2)$. [Liu, Majka, Szpruch, 2022]
- $F^\sigma = F + \frac{\sigma^2}{2} I(m)$. $d(m_1, m_2) = H(m_1|m_2)$.

# Descente de gradient entropie-Fisher

$d(m_1, m_2) = H(m_1|m_2)$, régularisé par $I$.
A chaque étape,

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

Calculs du premier ordre formels:

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

de sorte que

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

On espère que $m_{kh}^h \to m_t$ quand $h \to 0$ et $kh \to t$, où $m_t$ résout

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

# Conclusions

1. Problème d'optimisation champ-moyen avec régularisation de Fisher
2. Dynamique (MF Schrödinger, MF HJB, GD entropy-Fisher)
3. Convergence (pas de taux évident. Schrödinger : le spectre bouge... Nouvelles inégalités fonctionnelles ?)

Merci !