# Convex Optimization in the Space of Measure

Global vs Local Methods

Lénaïc Chizat*

2024

*EPFL, DOLA chair

# Convex optimization over measures

**Setting**

- $\mathcal{X}$ compact Riemannian $d$-manifold (torus, sphere), $d \geq 1$
- $\mathcal{M}(\mathcal{X})$ space of signed Borel measures on $\mathcal{X}$
- $\Phi : \mathcal{X} \to \mathbb{R}^n$ smooth filter/dictionnary, $y \in \mathbb{R}^n$ signal

$$F^* := \min_{\mu \in \mathcal{M}(\mathcal{X})} F(\mu), \qquad F(\mu) := \frac{1}{2}\Big\| \int \Phi(x)\, \mathrm{d}\mu(x) - y \Big\|_2^2 + \lambda \|\mu\|_{\mathsf{TV}}$$

Goal: given $\epsilon > 0$, find $\mu \in \mathcal{M}(\mathcal{X})$ such that $F(\mu) - F^* \leq \epsilon$

# Convex optimization over measures

**Setting**

- $\mathcal{X}$ compact Riemannian $d$-manifold (torus, sphere), $d \geq 1$
- $\mathcal{M}(\mathcal{X})$ space of signed Borel measures on $\mathcal{X}$
- $\Phi : \mathcal{X} \to \mathbb{R}^n$ smooth filter/dictionnary, $y \in \mathbb{R}^n$ signal

$$F^* := \min_{\mu \in \mathcal{M}(\mathcal{X})} F(\mu), \qquad F(\mu) := \frac{1}{2}\left\| \int \Phi(x)\, \mathrm{d}\mu(x) - y \right\|_2^2 + \lambda \|\mu\|_{\mathsf{TV}}$$

Goal: given $\epsilon > 0$, find $\mu \in \mathcal{M}(\mathcal{X})$ such that $F(\mu) - F^* \leq \epsilon$

**Global**

Time complexity in $\Theta(\epsilon^{-d})$.

- Frank-Wolfe
- **(Bregman) Gradient Descent**
- **Bilevel Mean-Field Langevin**

**Local (with non-degeneracy)**

Assuming $F(\mu_0) \leq F_0$, time complexity in $O(\log(1/\epsilon))$.

- "Sliding" particles with GD...
- ...**Wasserstein-Fisher-Rao GD**

Classification of some *algorithmic primitives*

# Outline

1. Global: Bregman Gradient Descent
2. Local: Wasserstein Fisher-Rao Gradient Descent
3. The Min-Max case, joint work Guillaume Wang

# Outline

1. Global: Bregman Gradient Descent
2. Local: Wasserstein Fisher-Rao Gradient Descent
3. The Min-Max case, joint work Guillaume Wang

# Convex optimization : $\infty$-dim analysis

**Approach**

Initialize *and fix* $(x_i)_{i=1}^m$ uniformly (on a grid/random) and solve the *convex* problem:

$$\min_{a\in\mathbb{R}^m} F_m(a), \qquad F_m(a) := \frac{1}{2}\Big\|\sum_{i=1}^m a_i\Phi(x_i) - y\Big\|_2^2 + \lambda\|a\|_1$$

# Convex optimization : $\infty$-dim analysis

**Approach**

Initialize *and fix* $(x_i)_{i=1}^m$ uniformly (on a grid/random) and solve the *convex* problem:

$$\min_{a\in\mathbb{R}^m} F_m(a), \qquad F_m(a) := \frac{1}{2}\Big\| \sum_{i=1}^m a_i \Phi(x_i) - y \Big\|_2^2 + \lambda \|a\|_1$$

**Infinite dimensional analysis**

- Classical guarantees explode as $m \to \infty$, non-informative
- In contrast, $\infty$-dimensional analysis leads to:
    - Classification of algorithms in terms of *cvge rates*;
    - Exhibits practical non-asymptotic cvge rates *before grid overfitting*

Fix $\tau \in \mathcal{P}(\mathcal{X})$ a reference measure and let $\mu = a\tau$ with $a \in L^1(\tau)$:

$$F^* := \min_{a\in L^1(\tau)} F(a), \quad F(a) := \frac{1}{2}\Big\| \int a(x)\Phi(x)\,\mathrm{d}\tau(x) - y \Big\|_2^2 + \lambda \|a\|_{L^1(\tau)}$$

# Bregman Proximal Gradient Methods

Setting: Minimize $F(a) = G(a) + H(a) =$ cvx smooth $+$ cvx proxable.
Power-entropy Bregman divergences, for $a, b \in L^1(\tau)$:

$$D_p(a,b) = \int \big(\eta(a) - \eta(b) - \eta'(b)(a-b)\big)\,\mathrm{d}\tau, \quad \eta(s) = \begin{cases} \frac{1}{p(p-1}s^p, & p \in ]1,2] \\ s\log(s) - s + 1, & p = 1 \end{cases}$$

**Proximal Gradient Method (PGM)**

Choose step-size $\eta > 0$ and initialization $a_1 \in \mathrm{dom}(H)$. For $k = 1, 2, \ldots$

$$a_{k+1} = \arg\min \langle a, G'[a_k]\rangle_{L^2(\tau)} + H(a) + \eta^{-1} D_p(a, a_k)$$

# Bregman Proximal Gradient Methods

Setting: Minimize $F(a) = G(a) + H(a) = $ cvx smooth + cvx proxable.

Power-entropy Bregman divergences, for $a, b \in L^1(\tau)$:

$$D_p(a,b) = \int \big(\eta(a) - \eta(b) - \eta'(b)(a-b)\big)\,\mathrm{d}\tau, \quad \eta(s) = \begin{cases} \frac{1}{p(p-1)} s^p, & p \in ]1,2] \\ s\log(s) - s + 1, & p = 1 \end{cases}$$

## Proximal Gradient Method (PGM)

Choose step-size $\eta > 0$ and initialization $a_1 \in \mathrm{dom}(H)$. For $k = 1, 2, \ldots$

$$a_{k+1} = \arg\min \langle a, G'[a_k]\rangle_{L^2(\tau)} + H(a) + \eta^{-1} D_p(a, a_k)$$

## Accelerated Proximal Gradient Method (APGM)

Choose step-size $\eta > 0$ and $a_1 \in \mathrm{dom}(H)$ and $\gamma_0 = 1$. For $k = 1, 2, \ldots$

1. $b_k = (1-\gamma_k)a_k + \gamma_k c_k$
2. $c_{k+1} = \arg\min \langle c, G'[b_k]\rangle_{L^2(\tau)} + H(c) + \eta^{-1} D_p(c, c_k)$
3. $a_{k+1} = (1-\gamma_k)a_k + \gamma_k c_{k+1}$
4. $\gamma_{k+1} = \frac{1}{2}\left(\sqrt{\gamma_k^4 + 4\gamma_k^2} - \gamma_k^2\right)$

# Our starting point: known guaranties

**Theorem[1] (adapted)**

For a small enough step-size $\eta$, if the iterates are bounded, it holds

$$F(a_k) - F(a) \leq \underbrace{\frac{4}{\eta k^{\beta}}}_{\xi_k} D_p(a, a_1), \quad \forall a \in L^1(\tau), \forall k \geq 1$$

where $\beta = 1$ for PGM and $\beta = 2$ for APGM.

- **Problem**: for sparse solutions $D_p(a^*, a_1) = \infty$ (in fact $a^* \notin L^1(\tau)$)
- **Workaround**: use instead[2,3]

$$F(a_k) - F^* \leq \inf_{a \in L^1(\tau)} \left(F(a) - F^*\right) + \xi_k D_p(a, a_1)$$

- Estimate the bound using a smoothing of $a^*$ as candidate

[1] Paul Tseng, (2010). *Approximation accuracy, gradient methods, and error bound for structured convex optimization.*

[2] Jacobs, Léger, Li, Osher (2018). *Solving large-scale optimization problems with a convergence rate*

[3] Chizat (2021). *Convergence Rates of Gradient Methods for Convex Optimization in the Space of Measures*

# Main results

**Theorem: convergence rates[1] [C. 2021]**

The convergence rate is $F(a_k) - F^* = O(\cdot)$ with $\cdot$ as follows:

| | PGM | APGM |
|---|---|---|
| $p = 1$ | $\log(k)k^{-1}$ | $\log(k)k^{-2}$ |
| $p > 1$ | $k^{-\frac{q}{(p-1)d+q}}$ | $k^{-\frac{2q}{(p-1)d+q}}$ |

**(a)** Convergence rates

| | $\Phi$ Lip. | $\nabla\Phi$ Lip. |
|---|---|---|
| $G'[\mu^*] > 0$ | $q = 1$ | $q = 2$ |
| $G'[\mu^*] = 0$ | $q = 2$ | $q = 4$ |

**(b)** Value of $q$ (highest that applies)

- rates are tight up to log factors (lower bounds on explicit instances)
- $G'[\mu^*] = 0$ means the penalization $H$ is inactive
- for $p = 2$ this is the rate of ISTA and FISTA (cursed!)
- for signed problems and $p = 1$: use hyperbolic entropy

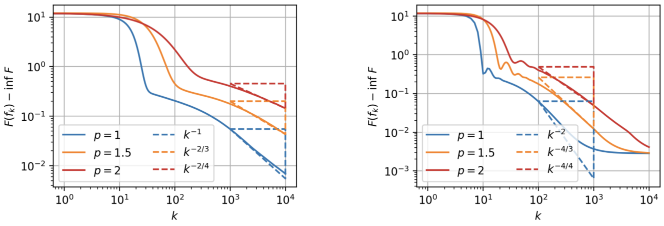
$$\eta(s) = s \cdot \mathrm{arcsinh}(s) - \sqrt{s^2 + 1} + 1$$

- Comparable to Frank-Wolfe, but *can be generically accelerated*

[1] Chizat (2021). *Convergence Rates of Gradient Methods for Convex Optimization in the Space of Measures.*

# Numerics

In pratice: these are the non-asymptotic rates (before overfitting the grid)




(a) PGM ($d = 2$, $q = 2$) (b) APGM ($d = 2$, $q = 2$)

Observed vs. theoretical rates on a non-degenerate sparse $2D$ deconvolution problem

$\rightsquigarrow$ $p = 1$ (APGM with hyperbolic entropy) is one order of magnitude faster than $p = 2$ (FISTA) on a large range of accuracies!

# Outline

1. Global: Bregman Gradient Descent
2. Local: Wasserstein Fisher-Rao Gradient Descent
3. The Min-Max case, joint work Guillaume Wang

# Re-parameterization with weighted particles

**Nonnegative case**

$$\min_{\mu \in \mathcal{M}_+(\mathcal{X})} F(\mu), \qquad F(\mu) := \frac{1}{2}\Big\| \int \Phi(\theta)\,\mathrm{d}\mu(\theta) - y \Big\|_2^2 + \lambda \mu(\mathcal{X})$$

**Particle formulation**

- Take $m \in \mathbb{N}$ particles with weight/position $(a_i, x_i) \in \mathbb{R}_+ \times \mathcal{X}$
- Parameterize $\mu_\theta = \frac{1}{m}\sum_{i=1}^m a_i \delta_{x_i}$ with $\theta = (a_i, x_i)_{i=1}^m$
- Find the minimizer (in $\theta$ and $m$) of

$$F_m(\theta) := \frac{1}{2}\Big(\frac{1}{m}\sum_{i=1}^m a_i \Phi(x_i) - y\Big)^2 + \frac{\lambda}{m}\sum_{i=1}^m a_i$$

$\rightsquigarrow$ convex in $(a_i)$, non-convex in $(x_i)$

# Re-parameterization with weighted particles

**Nonnegative case**

$$\min_{\mu\in\mathcal{M}_+(\mathcal{X})} F(\mu), \qquad F(\mu) := \frac{1}{2}\Big\| \int \Phi(\theta)\,\mathrm{d}\mu(\theta) - y\Big\|_2^2 + \lambda\mu(\mathcal{X})$$

**Particle formulation**

- Take $m \in \mathbb{N}$ particles with weight/position $(a_i, x_i) \in \mathbb{R}_+ \times \mathcal{X}$
- Parameterize $\mu_\theta = \frac{1}{m}\sum_{i=1}^m a_i\delta_{x_i}$ with $\theta = (a_i, x_i)_{i=1}^m$
- Find the minimizer (in $\theta$ and $m$) of

$$F_m(\theta) := \frac{1}{2}\Big(\frac{1}{m}\sum_{i=1}^m a_i\Phi(x_i) - y\Big)^2 + \frac{\lambda}{m}\sum_{i=1}^m a_i$$

$\rightsquigarrow$ convex in $(a_i)$, non-convex in $(x_i)$

**Signed case**

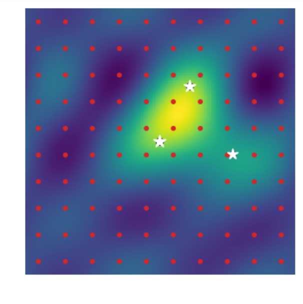
- Give sign to particles: $\theta = (a_i, x_i, \sigma_i)_{i=1}^m$ with $\sigma_i = \{+1, -1\}$.
- Equivalent to the unsigned case with $\tilde{\mathcal{X}}$ = two copies of $\mathcal{X}$

# Wasserstein-Fisher-Rao (aka Conic Particle) Gradient Flow

**Algorithm in continuous time (C. 2019)**

- Initialize with $(a_i(0), x_i(0))_{i=1}^m$ (potentially warm start)
- Compute $(\theta(t))_{t\geq 0}$ by following

$$\begin{cases} \frac{d}{dt} a_i(t) = -4m \cdot a_i(t) \nabla_{a_i} F_m(\theta(t)) \\ \frac{d}{dt} x_i(t) = -\frac{\alpha \cdot m}{a_i(t)} \nabla_{x_i} F_m(\theta(t)) \end{cases}$$



**Why multiplicative updates for weights?**

Initializing with $\theta(0) = (a_0, x_0)$

$\Leftrightarrow$

Initializing with $\theta(0) = ((a_0/2, x_0), (a_0/2, x_0))$

**Discrete time version (see paper)**

- Entropic mirror descent on $(a_i)$
- Gradient descent update on $(x_i)$

# Sparsity and optimality

**Assumption 1 (Uniqueness)**

There exists a unique minimizer which is sparse: $\mu^* = \sum_{i=1}^{m^*} a_i^* \delta_{x_i^*}$.

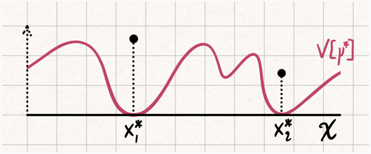
Let $V[\mu] \in \mathcal{C}^3(\mathcal{X})$ be the first variation of $F$ at $\mu$, characterized by

$$F(\mu + \epsilon\nu) = F(\mu) + \epsilon \int_{\mathcal{X}} V[\mu](x)\,\mathrm{d}\nu(x) + o(\epsilon), \quad \forall \nu \in \mathcal{M}(\mathcal{X}) \text{ admiss.}$$

**Proposition (Optimality conditions)**

The first variation of $F$ at $\mu^*$ satisfies

$$V[\mu^*] \geq 0 \quad \text{and} \quad \mathrm{spt}(\mu^*) = \{x_1^*, \dots, x_{m^*}^*\} \subset \{V[\mu^*] = 0\}.$$



# Kernels and Non-degeneracy assumption

**Definition (Interaction kernels)**

**Global interaction kernel** $K \in \mathcal{S}_+(m^*(d+1))$ (convention $\nabla_0 \phi = 2\phi$):

$$K_{(i,j),(i',j')} = \langle \sqrt{a_i^*}\nabla_j \phi(x_i^*, \cdot), \sqrt{a_{i'}^*}\nabla_{j'}\phi(x_{i'}^*, \cdot)\rangle_{L^2}$$

**Local interaction kernel** $H = \mathrm{diag}(H_i)_{i=1}^{m^*} \in \mathcal{S}_+(m^* d)$ with

$$H_i := \nabla^2 V[\mu^*](x_i^*)$$

**Definition (Non-degeneracy)**

We say that $F$ is **non-degenerate** iff:

- $K \succ 0$
- $\arg\min V[\mu^*] = \{x_1^*, \ldots, x_{m^*}^*\}$
- $H_i \succ 0$, $i \in \{1, \ldots, m^*\}$

$\rightsquigarrow$ Can be guaranteed a priori under spikes separation & noise level conditions[2,3]

[1] Duval and Peyré, 2015. *Exact Support Recovery for Sparse Spikes Deconvolution*

[2] Poon, Keriven, Peyré, 2018. *Support Localization [...]*

# Non-degeneracy vs. stability

**Wasserstein-Fisher-Rao metric**[1]

Define, for $\mu, \nu \in \mathcal{M}_+(\mathcal{X})$:

$$\mathrm{WFR}_2^2(\mu,\nu) := \min_{\gamma} \mathrm{KL}(\gamma_1|\mu) + \mathrm{KL}(\gamma_2|\nu) + \iint_{\mathcal{X}^2} c(x,y)\,\mathrm{d}\gamma(x,y)$$

where $\gamma \in \mathcal{M}_+(\mathcal{X}\times\mathcal{X})$ has marginals $\gamma_1, \gamma_2$ and $c(x,y) \approx \mathrm{dist}(x,y)^2/\alpha^2$

**Theorem[4] : quadratic growth (C., 2019)**

If $F$ is non-degenerate then $\exists F_0 > F^*$ such that if $F(\mu) \leq F_0$ then

$$\mathrm{WFR}_2^2(\mu,\mu^*) \lesssim F(\mu) - F^* \lesssim \mathrm{WFR}_2^2(\mu,\mu^*)$$

- All results are uniform in $m$ (hold even with $m = \infty$)
- $\mathrm{WFR}$ geometry appropriate for non-degenerate sparse problems

[1] Liero, Mielke, Savaré (2015). Kondratyev, Monsaingeon, Vorotnikov (2015). Chizat, Peyré, Schmitzer, Vialard (2015).

[2] Chizat (2019). *Sparse optimization on measures with over-parameterized gradient descent.*

# Back to dynamics

Rewriting $\mathrm{WFR}$ gradient flow using the first-variation $V$ gives:

$$\begin{cases} \frac{d}{dt} a_i(t) = -4a_i(t) V[\mu_t](x_i(t)) \\ \frac{d}{dt} x_i(t) = -\alpha \nabla V[\mu_t](x_i(t)) \end{cases}$$

where $\mu_t := \frac{1}{m} \sum_{i=1}^{m} a_i(t) \delta_{x_i(t)} \in \mathcal{M}_+(\mathcal{X})$.

**Proposition (Dynamics in the space of measures)**

The curve $(\mu_t)_t$ solves (distributionally) the PDE:

$$\partial_t \mu_t = \underbrace{\alpha \nabla \cdot \big(\mu_t \nabla V[\mu_t]\big)}_{\text{Drift}} - \underbrace{4\mu_t V[\mu_t]}_{\text{Reaction}}$$

This is the gradient flow of $F$ under the metric $\mathrm{WFR}$.

# Energy dissipation

**Energy dissipation** It holds $\frac{d}{dt}F(\mu_t) = -\|\nabla_{\mathrm{WFR}} F(\mu_t)\|^2$ with the squared-norm of $\mathrm{WFR}$ gradient:

$$\|\nabla_{\mathrm{WFR}} F(\mu)\|^2 := \int_{\mathcal{X}} \left(\alpha \|\nabla V[\mu](x)\|^2 + 4|V[\mu](x)|^2\right) \mathrm{d}\mu(x)$$

**Theorem : PL inequality[1] C. 2019)**

If $F$ is non-degenerate then $\exists F_0 > F^*$ such that if $F(\mu) < F_0$ then

$$\|\nabla_{\mathrm{WFR}} F[\mu]\|^2 \gtrsim F(\mu) - F^*$$

**Corollary**

If $F$ is non-degenerate then $\exists C_0, C_1 > 0$ independent of $m$ such that

$$F(\mu_0) - F^* \leq C_0 \quad \Rightarrow \quad F(\mu_t) - F^* \leq C_0 e^{-C_1 t}.$$

- Overall time complexity $O(m^2 n \log(1/\epsilon))$
- PL inequality and growth are related *in finite dimension*[2]

---

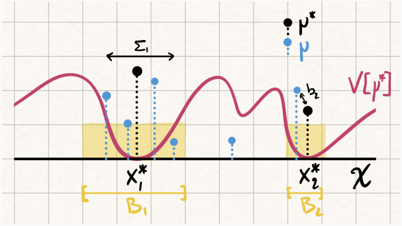
[1] Chizat (2019). *Sparse optimization on measures with over-parameterized gradient descent.*

[2] Rebjock, Boumal (2023). *Fast convergence to non-isolated minima [...]*

# Proof idea and local expansion

Decompose $\mu$ into local moments in small balls $B_i$ around each $x_i^*$:

- local biases $b_i \in \mathbb{R}^{d+1}$
- local covariances $\Sigma_i \in \mathbb{R}^{d\times d}$




**Local Taylor expansion of $F$ around $\mu^*$**

$$F(\mu) - F^* \approx \underbrace{\frac{1}{2} b^\intercal (K+H) b}_{\text{Bias term (local+global)}} + \underbrace{\sum_{i=1}^{m^*} a_i \operatorname{tr}(\Sigma_i H_i)}_{\text{Variance term (local)}} + \underbrace{\int_{\mathcal{X}\setminus(\cup B_i)} V[\mu^*]\,\mathrm{d}\mu}_{\text{Mass sent to 0}}$$

# Outline

# Bilinear minmax problem

Continuous strategy spaces $\mathcal{X}, \mathcal{Y}$, pay-off function $f \in \mathcal{C}^3(\mathcal{X} \times \mathcal{Y})$.

**Mixed Nash equilibrium of continuous games**

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu \in \mathcal{P}(\mathcal{Y})} \left\{ F(\mu, \nu) := \int_{\mathcal{X} \times \mathcal{Y}} f(x, y) \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y) \right\}$$

# Bilinear minmax problem

Continuous strategy spaces $\mathcal{X}, \mathcal{Y}$, pay-off function $f \in \mathcal{C}^3(\mathcal{X} \times \mathcal{Y})$.

**Mixed Nash equilibrium of continuous games**

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \max_{\nu \in \mathcal{P}(\mathcal{Y})} \left\{ F(\mu, \nu) := \int_{\mathcal{X} \times \mathcal{Y}} f(x, y) \, \mathrm{d}\mu(x) \, \mathrm{d}\nu(y) \right\}$$

**Why bilinear?**

- Applications to robust training of 2-layer Neural Networks
- Contains all difficulties of general convex-concave objectives
- In particular, explicit (or continuous-time) fixed-grid methods do *not* converge for bilinear games $\rightsquigarrow$ need implicit steps

Parameterize $\mu = \sum_{i=1}^m a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^m b_i \delta_{y_i}$ with $a, b \in \Delta^{m-1}$

# Conic Particle Proximal Point (CPPP)

$$(\theta^\mu_{k+1}, \theta^\nu_{k+1}) = \underset{\theta^\mu=(a_i,x_i)_{i=1}^m}{\arg\min}\ \mathrm{argmax}_{\theta^\nu=(b_i,y_i)_{i=1}^m} F_m(\theta^\mu, \theta^\nu)$$
$$+ \frac{1}{\eta}\mathrm{KL}(a|a(k)) + \frac{1}{\alpha\eta}\sum_i a_i(k)\mathrm{dist}(x_i, x_i(k))^2$$
$$+ \frac{1}{\eta}\mathrm{KL}(b|b(k)) + \frac{1}{\alpha\eta}\sum_i b_i(k)\mathrm{dist}(y_i, y_i(k))^2$$

[1] Wang, Chizat (2022). *An Exponentially Converging Particle Method for Mixed Nash Equilibria [...].*

# Conic Particle Proximal Point (CPPP)

$$(\theta^\mu_{k+1}, \theta^\nu_{k+1}) = \underset{\theta^\mu=(a_i,x_i)_{i=1}^m}{\arg\min}\ \operatorname{argmax}_{\theta^\nu=(b_i,y_i)_{i=1}^m} F_m(\theta^\mu, \theta^\nu)$$
$$+ \frac{1}{\eta}\mathrm{KL}(a|a(k)) + \frac{1}{\alpha\eta}\sum_i a_i(k)\mathrm{dist}(x_i, x_i(k))^2$$
$$+ \frac{1}{\eta}\mathrm{KL}(b|b(k)) + \frac{1}{\alpha\eta}\sum_i b_i(k)\mathrm{dist}(y_i, y_i(k))^2$$

**Theorem : Local convergence**[1]

If $F$ admits a unique *sparse*, non-degenerate saddle $(\mu^*, \nu^*)$ and given $\alpha > 0$, there exists $C_i > 0$ (independent of $m$) such that

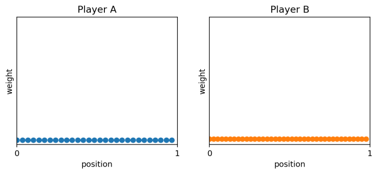
$$\eta < C_0,\ \ \Delta(\theta^\mu(0), \theta^\nu(0)) \le C_1 \quad \Rightarrow \quad \Delta(\theta^\mu(t), \theta^\nu(t)) \le C_2 e^{-C_3\eta^2 t}$$

where $\Delta$ denotes the primal-dual gap.

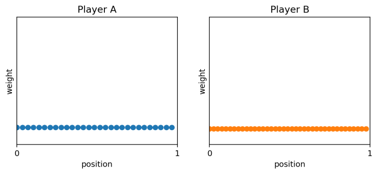
$\rightsquigarrow$ In practice: replace proximal point by extra-gradient
$\rightsquigarrow$ Continuous-time does not converge *in general*

[1] Wang, Chizat (2022). *An Exponentially Converging Particle Method for Mixed Nash Equilibria [...].*

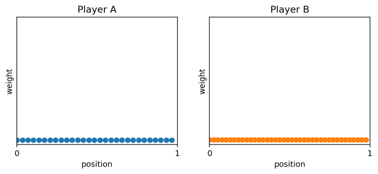
# Surprising behavior



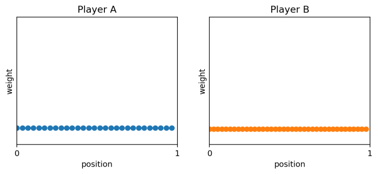

Implicit, fixed positions




Implicit, everything moves




Explicit, fixed positions




Explicit, everything moves

# Local analysis

The ODE $z'(t) = M(z(t) - z^*)$ satisfies $\|z(t) - z^*\|_2 = \tilde{\Theta}(e^{t \cdot \mathrm{sa}(M)})$.

**Spectral Abscissa (local convergence rate)**

$$\mathrm{sa}(M) := \max_{\lambda \in \mathrm{Spectrum}(M)} \mathrm{Real}(\lambda) \leq 0$$

# Local analysis

The ODE $z'(t) = M(z(t) - z^*)$ satisfies $\|z(t) - z^*\|_2 = \tilde{\Theta}(e^{t \cdot \mathrm{sa}(M)})$.

**Spectral Abscissa (local convergence rate)**

$$\mathrm{sa}(M) := \max_{\lambda \in \mathrm{Spectrum}(M)} \mathrm{Real}(\lambda) \leq 0$$

Generic game $f(x, y) \in \mathcal{C}^3(\mathbb{R}^d \times \mathbb{R}^d)$ with saddle $z^* = (x^*, y^*)$.

$$M = \begin{bmatrix} -\nabla^2_{xx} f & -\nabla^2_{xy} f \\ \nabla^2_{xy} f^\top & \nabla^2_{yy} f \end{bmatrix} (z^*) = -\underbrace{\begin{bmatrix} \nabla^2_{xx} f & 0 \\ 0 & \nabla^2_{yy} f \end{bmatrix}}_{S \text{ (psd)}} + \underbrace{\begin{bmatrix} 0 & -\nabla^2_{xy} f \\ \nabla^2_{xy} f^\top & 0 \end{bmatrix}}_{A \text{ (antisym.)}}$$

# Local analysis

The ODE $z'(t) = M(z(t) - z^*)$ satisfies $\|z(t) - z^*\|_2 = \tilde{\Theta}(e^{t \cdot \mathrm{sa}(M)})$.

**Spectral Abscissa (local convergence rate)**

$$\mathrm{sa}(M) := \max_{\lambda \in \mathrm{Spectrum}(M)} \mathrm{Real}(\lambda) \leq 0$$

Generic game $f(x, y) \in \mathcal{C}^3(\mathbb{R}^d \times \mathbb{R}^d)$ with saddle $z^* = (x^*, y^*)$.

$$M = \begin{bmatrix} -\nabla^2_{xx} f & -\nabla^2_{xy} f \\ \nabla^2_{xy} f^\top & \nabla^2_{yy} f \end{bmatrix}(z^*) = -\underbrace{\begin{bmatrix} \nabla^2_{xx} f & 0 \\ 0 & \nabla^2_{yy} f \end{bmatrix}}_{S \text{ (psd)}} + \underbrace{\begin{bmatrix} 0 & -\nabla^2_{xy} f \\ \nabla^2_{xy} f^\top & 0 \end{bmatrix}}_{A \text{ (antisym.)}}$$

- If $A = 0$, then $\mathrm{sa}(M) = -\min \mathrm{EigenVal}(S)$
- If $f$ is $\mu$-strongly convex/concave, then $\mathrm{sa}(M) \leq -\mu$
- in bilinear games, $S = 0$ and $\mathrm{sa}(M) = 0$

$\rightsquigarrow$ **What is $\mathrm{sa}(M)$ in general?**

# Local convergence rate

**A remark (Partial curvature suffices)**

Let $S \neq 0$ a psd matrix. Then for almost any $A$, $\mathrm{sa}(M) < 0$.

[1] Wang, Chizat (2023). *Local Convergence of Gradient Methods for Min-Max Games under Partial Curvature.*

# Local convergence rate

**A remark (Partial curvature suffices)**

Let $S \neq 0$ a psd matrix. Then for almost any $A$, $\mathrm{sa}(M) < 0$.

**Theorem: Mean curvature matters, not min curvature**[1]

Let $S$ a psd matrix. If $\nabla^2_{xy} f(z^*)$ (the interaction part) is nonsingular, has distinct eigenvalues and its singular vectors are uniformly distributed, then for $M_\alpha = A + \alpha S$ it holds

$$\mathbf{E}\big[\mathrm{sa}(M_\alpha)\big] = \underbrace{-\frac{\alpha\,\mathrm{tr}(S)}{d}}_{\text{Average of eigenvalues}} + O(\alpha^3) + \alpha\epsilon(d)$$

where $|\epsilon(d)| \leq 2\sqrt{\log(d)}/(\mathrm{tr}(S)/\|S\|_F)$ is small if $S$'s spectrum is not sparse.

$\rightsquigarrow$ In the 'particle' case, the convergence rate is generically $\Theta(\alpha^2\eta)$

[1] Wang, Chizat (2023). *Local Convergence of Gradient Methods for Min-Max Games under Partial Curvature.*

# Conclusion

### Last remarks

- Theory suggests the following building blocks:
  - Global: APGM with hyperbolic entropy
  - Local: WFR gradient descent (can be accelerated as well)
- Not discussed:
  - Role of noise and *Mean-Field Langevin* dynamics
  - Using a single algorithm instead of two

### Based on the following papers:


I C. (2021). *Convergence Rates of Gradient Methods for Convex Optimization in the Space of Measures.*

II C. (2019). *Sparse Optimization on Measures with Over-parameterized Gradient Descent.*

III Wang, C. (2022). *An Exponentially Converging Particle Method for Mixed Nash Equilibria [...]*. and Wang, C. (2023). *Local Convergence of Gradient Methods for Min-Max Games under Partial Curvature*