# A Distributed Algorithm for Wasserstein Proximal Operator Splitting

Abhishek Halder

Department of Applied Mathematics
University of California, Santa Cruz
Santa Cruz, CA 95064

Joint work with I. Nodozi, A.M. Teter (UC Santa Cruz)

Mechanical Engineering Seminar Series, University of Alabama, Tuscaloosa
November 04, 2022

# Topic of this talk

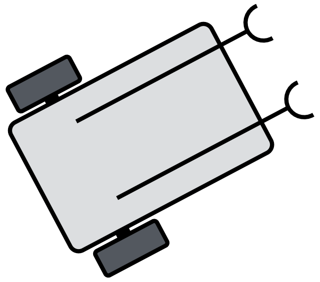
**Optimization over the space of measures a.k.a. distributions**

# What do we mean by measure a.k.a. distribution

**measure a.k.a. distribution = mass**

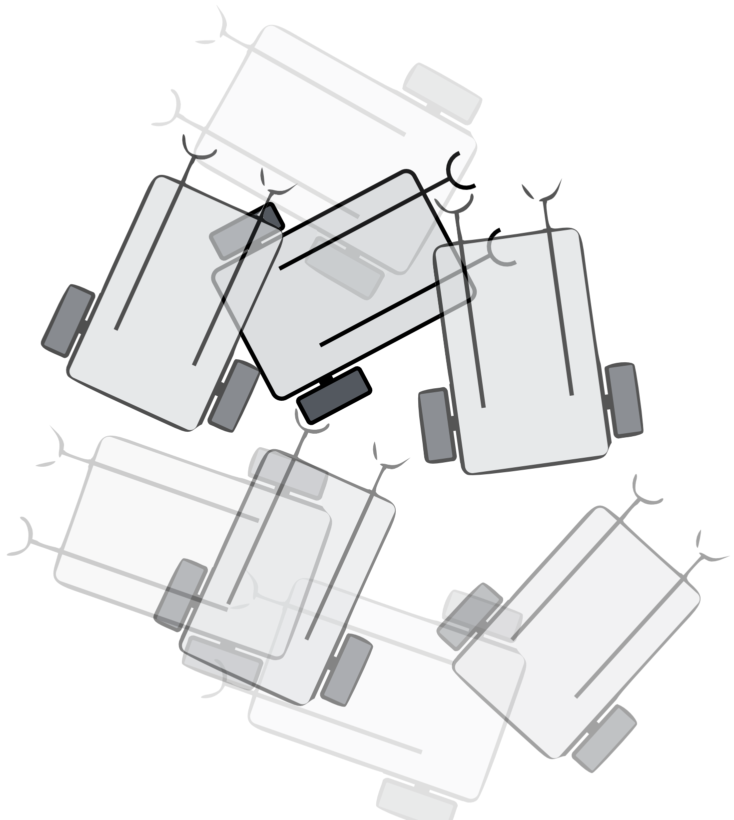
**mass = density $\times$ volume**

# Probability Distribution



$$\boldsymbol{x}(t) = \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

# Probability Distribution

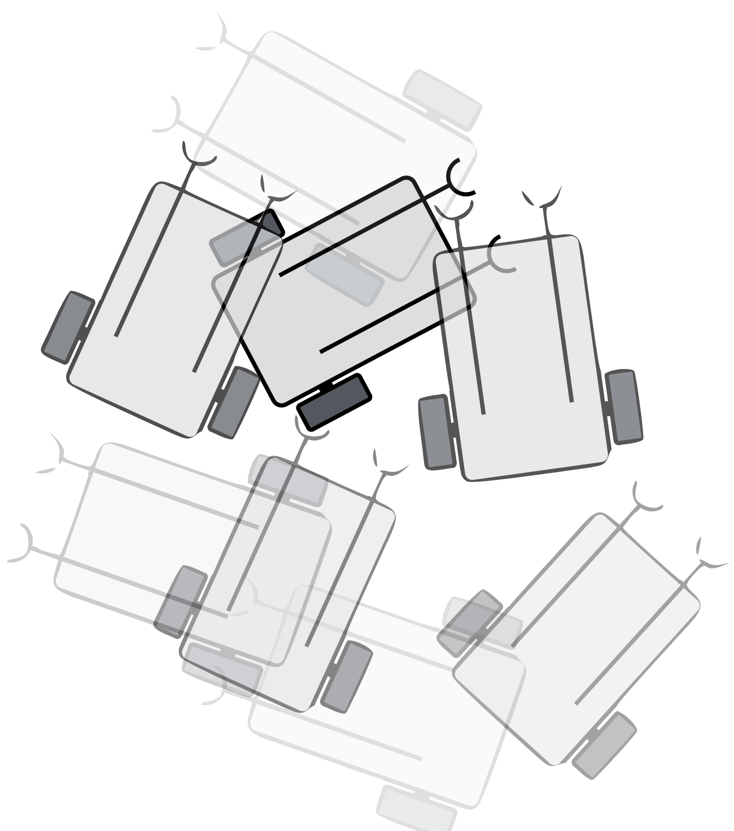


$$\boldsymbol{x}(t) = \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

$$\rho\left(\boldsymbol{x}, t\right) : \mathcal{X} \times [0, \infty) \mapsto \mathbb{R}_{\geq 0}$$

$$\int_{\mathcal{X}} \mathrm{d}\mu = \int_{\mathcal{X}} \rho \, \mathrm{d}\boldsymbol{x} = 1 \quad \text{for all } t \in [0, \infty)$$

# Probability Distribution



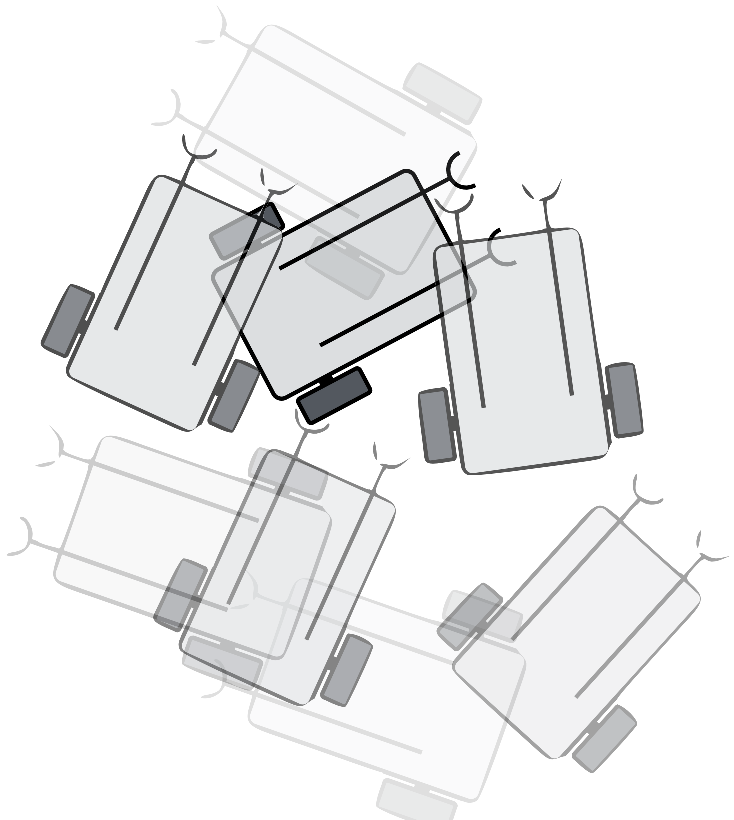
$$\boldsymbol{x}(t) = \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

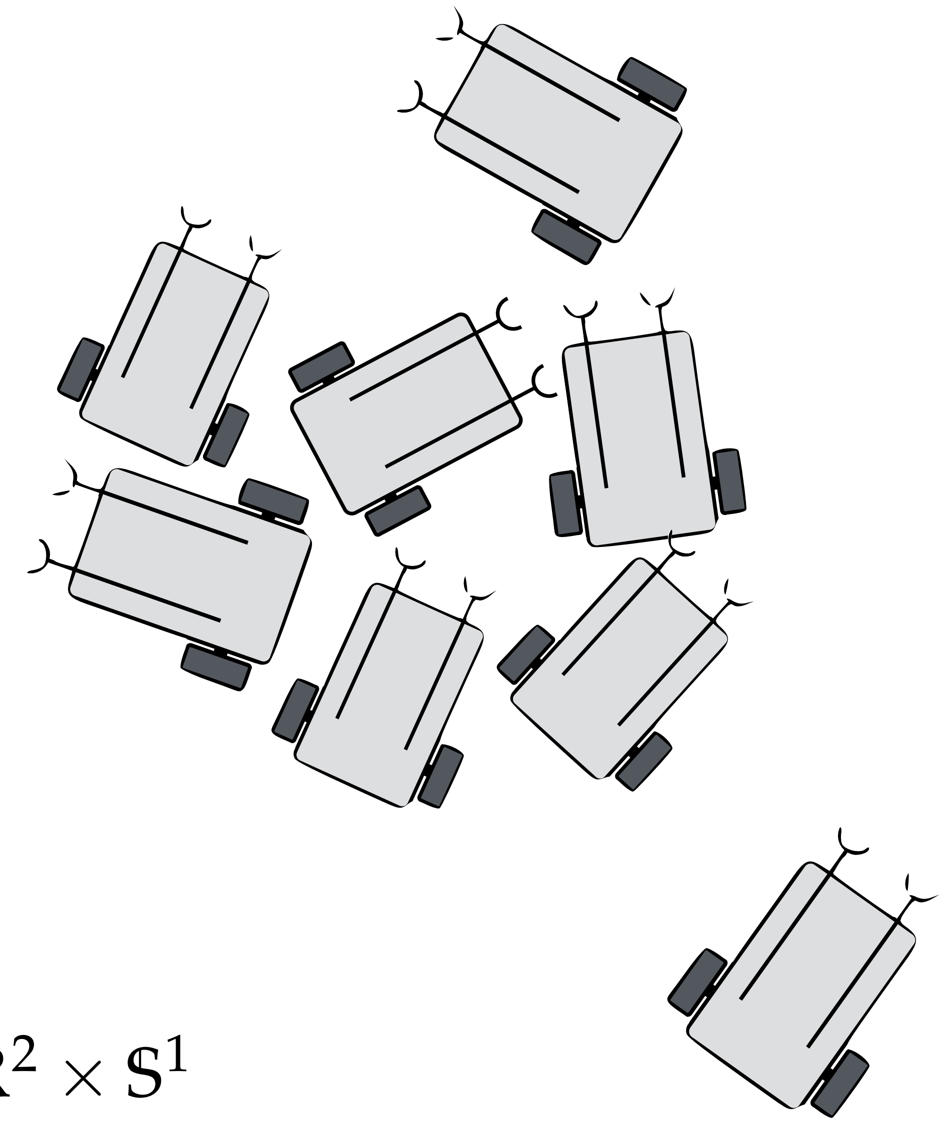
$$\rho\left(\boldsymbol{x}, t\right) : \mathcal{X} \times [0, \infty) \mapsto \mathbb{R}_{\geq 0}$$

probability measure probability density function

$$\int_{\mathcal{X}} \mathrm{d}\mu = \int_{\mathcal{X}} \rho \, \mathrm{d}\boldsymbol{x} = 1 \quad \text{for all } t \in [0, \infty)$$

# Probability Distribution Population Distribution





$$\boldsymbol{x}(t) = \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

$$\rho\left(\boldsymbol{x}, t\right) : \mathcal{X} \times [0, \infty) \mapsto \mathbb{R}_{\geq 0}$$

population measure population density function

$$\int_{\mathcal{X}} \mathrm{d}\mu = \int_{\mathcal{X}} \rho \, \mathrm{d}\boldsymbol{x} = 1 \quad \text{for all } t \in [0, \infty)$$

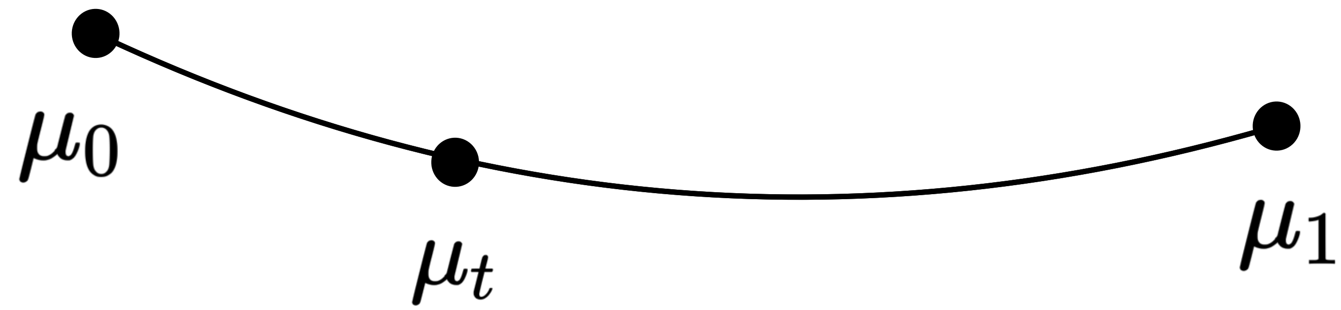
# Geometry on the Space of Prob. Measures


2-Wasserstein distance metric

$$W_2 := \left( \inf_{\mu, \boldsymbol{v}} \left\{ \frac{1}{2} \int_0^1 \int_{\mathcal{X}} \|\boldsymbol{v}\|^2 \mathrm{d}\mu \, \mathrm{d}t \right\} \right)^{1/2}$$

$$\text{subject to} \quad \frac{\partial \mu}{\partial t} = -\nabla \cdot (\mu \boldsymbol{v}), \; \mu(t=0,\cdot) = \mu_0, \; \mu(t=1,\cdot) = \mu_1$$

# Geometry on the Space of Prob. Measures




2-Wasserstein distance metric

$$W_2 := \left( \inf_{\mu, \boldsymbol{v}} \left\{ \frac{1}{2} \int_0^1 \int_{\mathcal{X}} \|\boldsymbol{v}\|^2 \mathrm{d}\mu \, \mathrm{d}t \right\} \right)^{1/2}$$

$$\text{subject to} \quad \frac{\partial \mu}{\partial t} = -\nabla \cdot (\mu \boldsymbol{v}), \;\; \mu(t=0,\cdot) = \mu_0, \;\; \mu(t=1,\cdot) = \mu_1$$

Measure-valued geodesic path for any $t \in [0,1]$

$$\mu_t = \underset{\nu \in \mathcal{P}_2(\mathcal{X})}{\arg\inf} \left\{ (1-t) W_2^2(\mu_0, \nu) + t W_2^2(\mu_1, \nu) \right\}$$

manifold of probability measures supported on $\mathcal{X}$ with finite second moments

# Measure-valued Optimization Problems

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; F(\mu)$$

2-Wasserstein geodescially convex functional

Space of Borel probability measures on $\mathbb{R}^d$ with finite second moments

In many applications, we have additive structure:

$$F(\mu) = F_1(\mu) + F_2(\mu) + \ldots + F_n(\mu)$$

where each $F_i : \mathscr{P}_2\left(\mathbb{R}^d\right) \mapsto (-\infty, +\infty]$ is proper, lsc, and 2-Wasserstein geodescially convex

# Connection with Wasserstein Gradient Flows

$$\frac{\partial \mu}{\partial t} = -\nabla^{W_2} F(\mu) := \nabla \cdot \left( \mu \nabla \frac{\delta F}{\delta \mu} \right) \qquad (\star)$$

Wasserstein gradient

Minimizer of $\arg\inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$ $\leftsquigarrow$ Stationary solution of $(\star)$

Transient solution of $(\star)$ $\rightsquigarrow$ Discrete time-stepping realizing grad. descent of $\arg\inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F(\mu)$

# Connection with Wasserstein Gradient Flows

$$\frac{\partial \mu}{\partial t} = -\nabla^{W_2} F(\mu) := \nabla \cdot \left( \mu \nabla \frac{\delta F}{\delta \mu} \right) \qquad (\star)$$

Wasserstein gradient

Minimizer of $\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; F(\mu)$ $\leftsquigarrow$ Stationary solution of $(\star)$

Transient solution of $(\star)$ $\rightsquigarrow$ Discrete time-stepping realizing grad. descent of $\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; F(\mu)$

Wasserstein proximal recursion à la Jordan-Kinderlehrer-Otto (JKO) scheme

# Gradient Flows

### Gradient Flow in $\mathcal{X}$

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = -\nabla f(\boldsymbol{x}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0$$

**Recursion:**

$$\begin{aligned}\boldsymbol{x}_k &= \boldsymbol{x}_{k-1} - h\nabla f(\boldsymbol{x}_k)\\ &= \underset{\boldsymbol{x}\in\mathcal{X}}{\arg\min}\left\{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_{k-1}\|_2^2 + hf(\boldsymbol{x})\right\}\\ &=: \mathrm{prox}_{hf}^{\|\cdot\|_2}(\boldsymbol{x}_{k-1})\end{aligned}$$

**Convergence:**

$$\boldsymbol{x}_k \to \boldsymbol{x}(t = kh) \quad \text{as} \quad h \downarrow 0$$

**$f$ as Lyapunov function:**

$$\frac{\mathrm{d}}{\mathrm{d}t} f = -\|\nabla f\|_2^2 \leq 0$$

### Gradient Flow in $\mathcal{P}_2(\mathcal{X})$

$$\frac{\partial \mu}{\partial t} = -\nabla^W F(\mu), \quad \mu(\boldsymbol{x}, 0) = \mu_0$$

**Recursion:**

$$\begin{aligned}\mu_k &= \mu(\cdot, t = kh)\\ &= \underset{\mu\in\mathcal{P}_2(\mathcal{X})}{\arg\min}\left\{\frac{1}{2}W^2(\mu,\mu_{k-1}) + hF(\mu)\right\}\\ &=: \mathrm{prox}_{hF}^{W}(\mu_{k-1})\end{aligned}$$

**Convergence:**

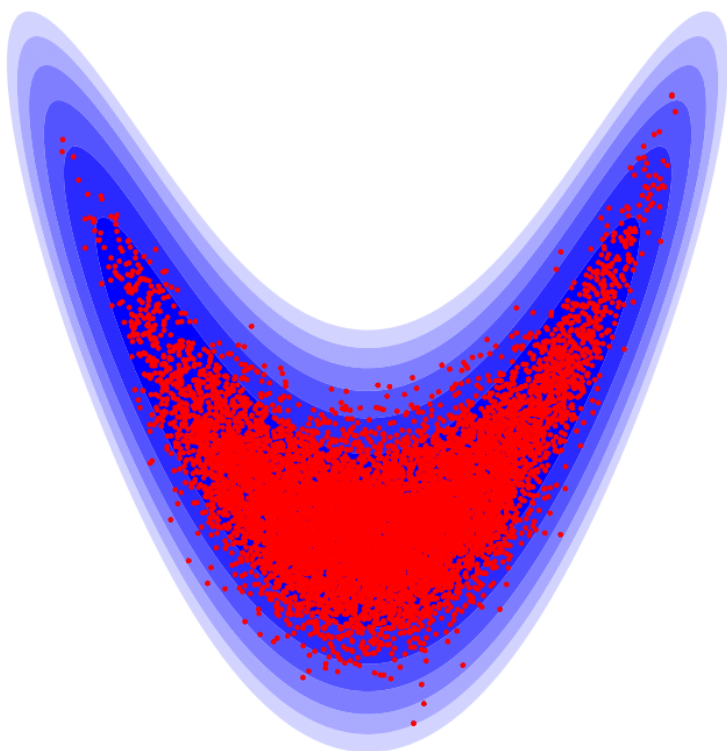
$$\mu_k \to \mu(\cdot, t = kh) \quad \text{as} \quad h \downarrow 0$$

**$F$ as Lyapunov functional:**

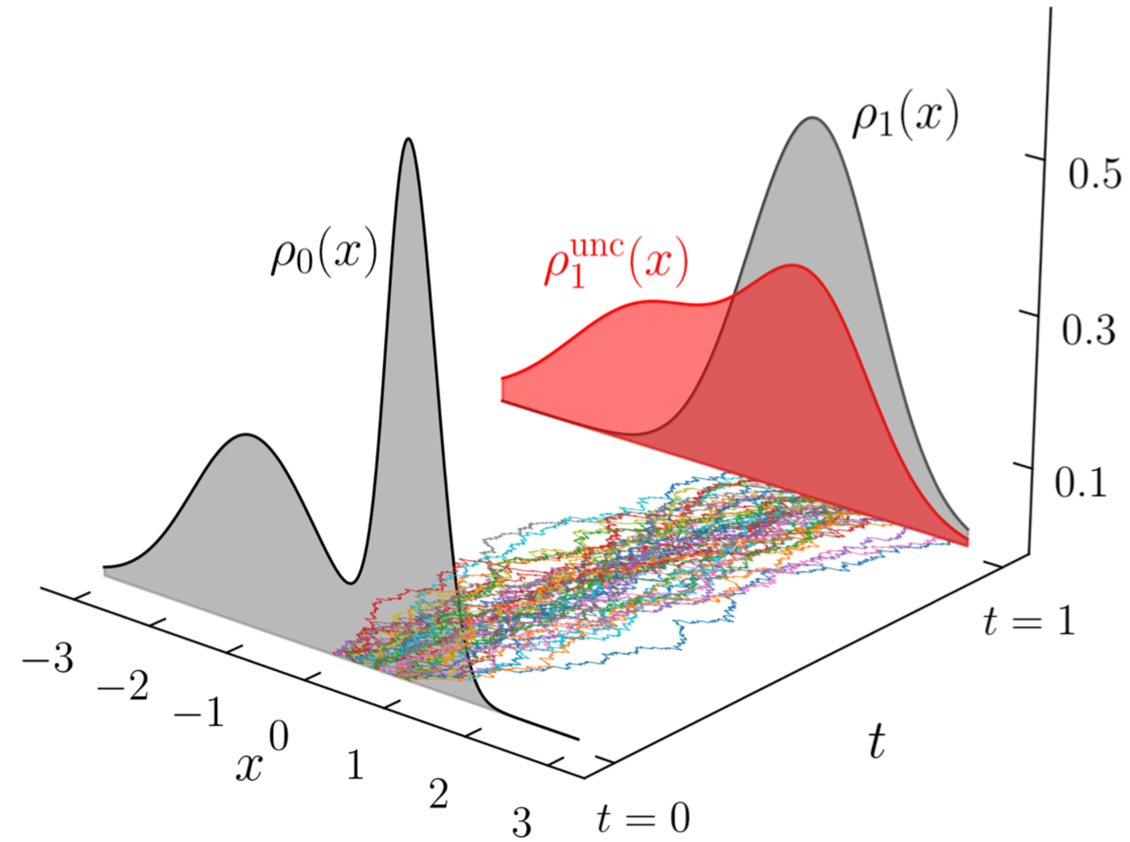
$$\frac{\mathrm{d}}{\mathrm{d}t} F = -\mathbb{E}_\mu\left[\left\|\nabla\frac{\delta F}{\delta \mu}\right\|_2^2\right] \leq 0$$

# Motivating Applications

Langevin sampling from an unnormalized prior



Optimal control of distributions a.k.a. Schrödinger bridge problems





Stramer and Tweedie, *Methodology and Computing in Applied Probability*, 1999

Jarner and Hansen, *Stochastic Processes and their Applications*, 2000

Roberts and Stramer, *Methodology and Computing in Applied Probability*, 2002

Vempala and Wibisino, *NeurIPS*, 2019

Chen, Georgiou and Pavon, *SIAM Review*, 2021

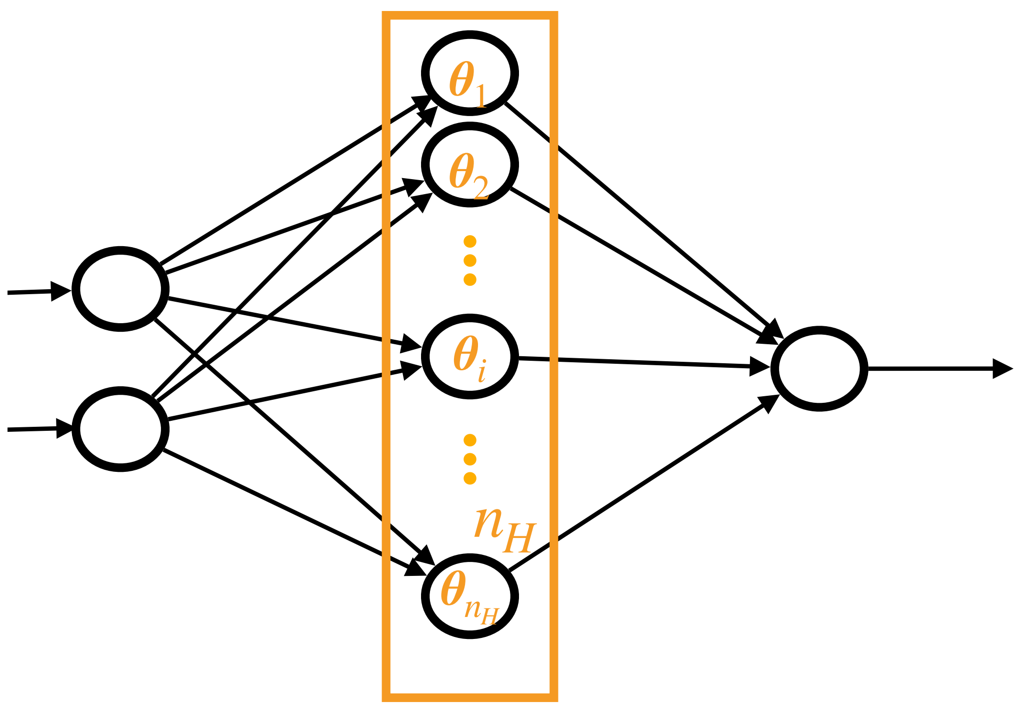
Chen, Georgiou and Pavon, *SIAM Journal on Applied Mathematics*, 2016

Chen, Georgiou and Pavon, *Journal on Optimization Theory and Applications*, 2016

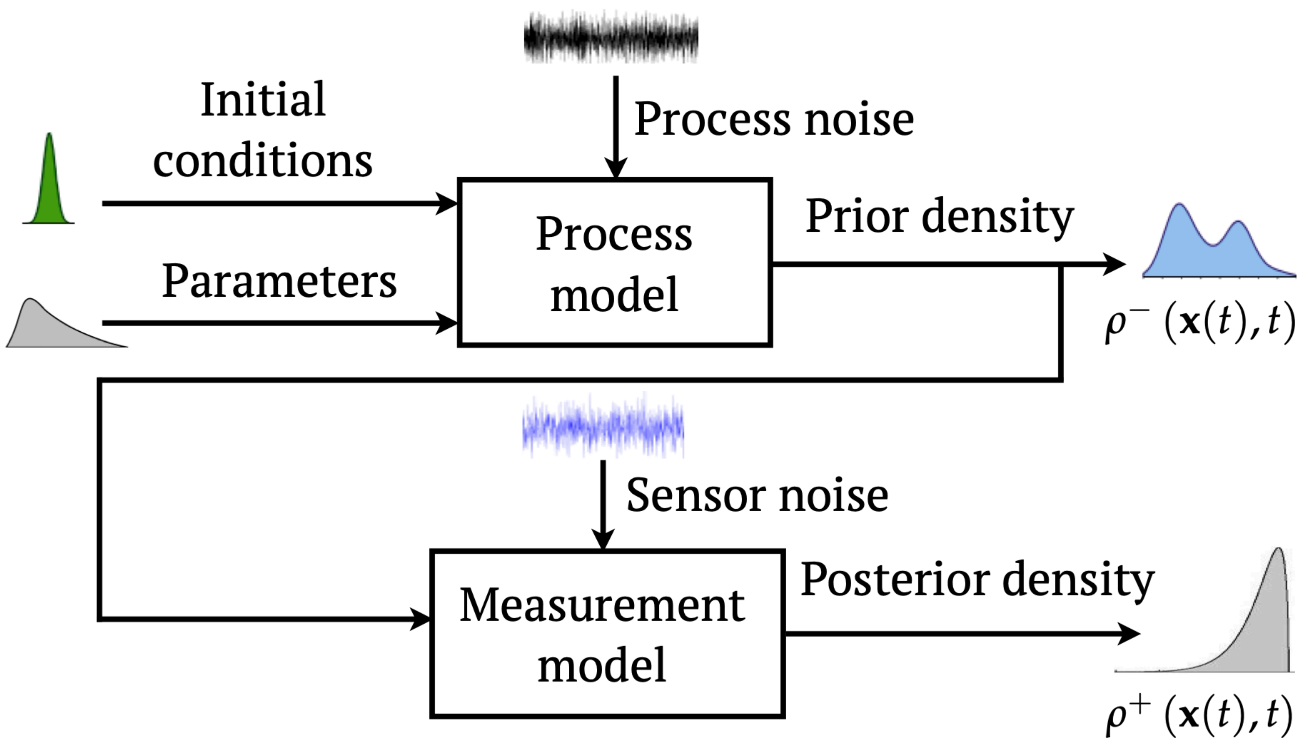
Caluya and Halder, *IEEE Transactions on Automatic Control*, 2021

# Motivating Applications (contd.)

Mean field learning dynamics in neural networks




Prediction and estimation of time-varying joint state probability densities




Mei, Montanari and Nguyen, *Proceedings of the National Academy of Sciences*, 2018

Chizat and Bach, *NeurIPS*, 2018

Rotskoff and Vanden-Eijnden, *NeurIPS*, 2018

Sirignano and Spiliopoulos, *Stochastic Processes and their Applications*, 2020

Caluya and Halder, *IEEE Transactions on Automatic Control*, 2019

Halder and Georgiou, *CDC*, 2019

Halder and Georgiou, ACC, 2018

Halder and Georgiou, *CDC*, 2017

# Many Recently Proposed Algorithms to Solve Measure-valued Optimization Problems

Peyré, *SIAM Journal on Imaging Sciences*, 2015

Benamou, Carlier and Laborde, *ESAIM: Proceedings and Surveys*, 2016

Carlier, Duval, Peyré and Schimtzer, *SIAM Journal on Mathematical Analysis*, 2017

Karlsson and Ringh, *SIAM Journal on Imaging Sciences*, 2017

Caluya and Halder, *IEEE Transactions on Automatic Control*, 2019

Carrillo, Craig, Wang and Wei, *Foundations of Computational Mathematics*, 2021

Mokrov, Korotin, Li, Gnevay, Solomon, and Burnaev, *NeurIPS*, 2021

Alvarez-Melis, Schiff, and Mroueh, *NeurIPS*, 2021

# Many Recently Proposed Algorithms to Solve Measure-valued Optimization Problems

Peyré, *SIAM Journal on Imaging Sciences*, 2015

Benamou, Carlier and Laborde, *ESAIM: Proceedings and Surveys*, 2016

Carlier, Duval, Peyré and Schimtzer, *SIAM Journal on Mathematical Analysis*, 2017

Karlsson and Ringh, *SIAM Journal on Imaging Sciences*, 2017

Caluya and Halder, *IEEE Transactions on Automatic Control*, 2019

Carrillo, Craig, Wang and Wei, *Foundations of Computational Mathematics*, 2021

Mokrov, Korotin, Li, Gnevay, Solomon, and Burnaev, *NeurIPS*, 2021

Alvarez-Melis, Schiff, and Mroueh, *NeurIPS*, 2021

But all require centralized computing

# Centralized Computing Case Study: Mean Field SGD Dynamics in NN Classification

Free energy functional: $F(\mu) = R\Big(\hat{f}\,(\boldsymbol{x}, \mu)\Big)$

For quadratic loss:

$$F(\mu) = F_0 + \int_{\mathbb{R}^p} V(\boldsymbol{\theta})\mathrm{d}\mu(\boldsymbol{\theta}) + \int_{\mathbb{R}^{2p}} U(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}})\mathrm{d}\mu(\boldsymbol{\theta})\mathrm{d}\mu(\tilde{\boldsymbol{\theta}})$$

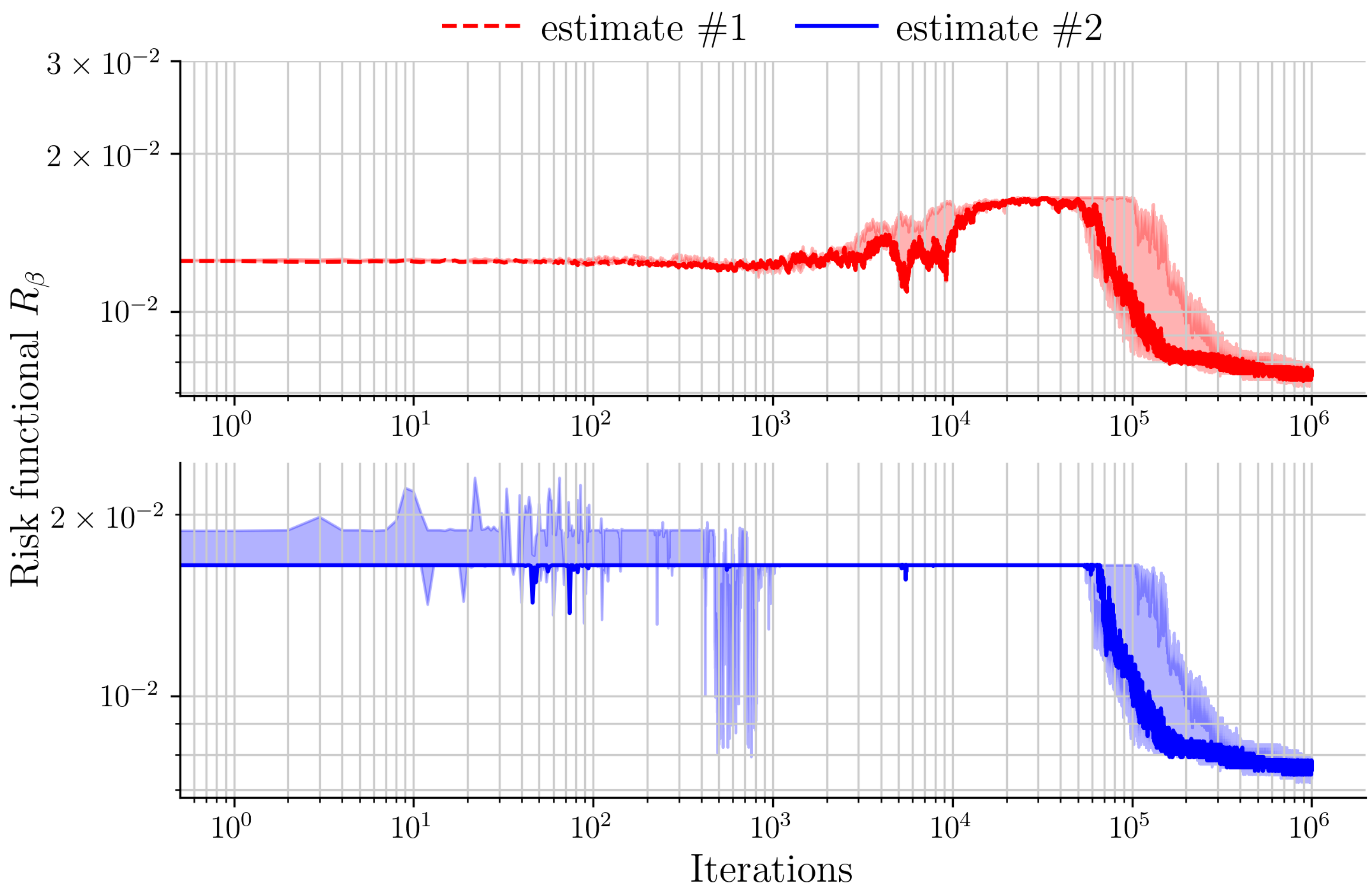
depend on activation functions of the NN

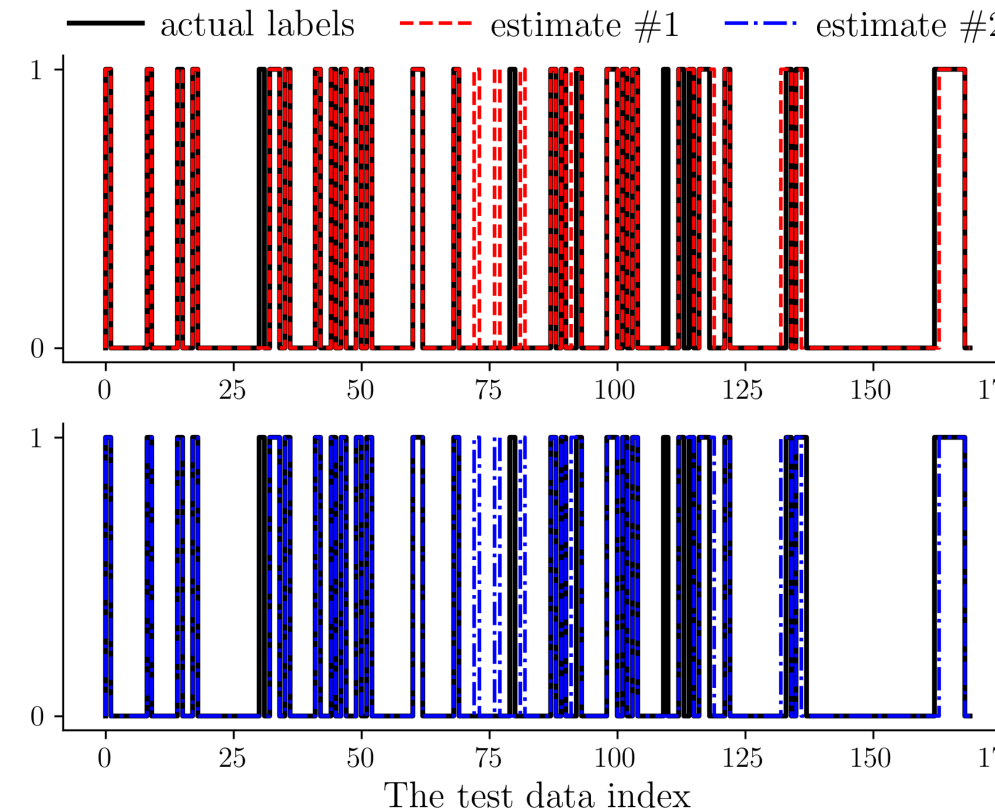
Neuronal population measure dynamics: $\dfrac{\partial \mu}{\partial t} = \nabla \cdot \left(\mu \nabla \dfrac{\delta F}{\delta \mu}\right) =: -\nabla^{W_2} F(\mu)$

Wasserstein proximal recursion: $\mu_{k+1} = \mathrm{prox}_{hF}^{W}(\mu_k)$

# Centralized Computing Case Study: Mean Field SGD Dynamics in NN Classification

## Case study: Wisconsin Breast Cancer (Diagnostic) Data Set




Classification accuracy for the WBDC dataset

| $\beta$ | Estimate #1 | Estimate #2 |
|---|---|---|
| 0.03 | 91.17% | 92.35% |
| 0.05 | 92.94% | 92.94% |
| 0.07 | 78.23% | 92.94% |

**CPU:** 3.4 GHz 6 core intel i5 8GB RAM (≈ 33 hrs runtime)

**GPU:** Jetson TX2 NVIDIA Pascal GPU 256 CUDA cores, 64 bit NVIDIA Denver + ARM Cortex A57 CPUs (≈ 2 hrs runtime)

# Present Work: Distributed Algorithm

$$\operatorname*{arg\,inf}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \; F_1(\mu) + F_2(\mu) + \ldots + F_n(\mu)$$

# Present Work: Distributed Algorithm

$$\operatorname*{arg\,inf}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \; F_1(\mu) + F_2(\mu) + \ldots + F_n(\mu)$$

**Main idea:** re-write

$$\begin{aligned} \operatorname*{arg\,inf}_{(\mu_1,\ldots,\mu_n,\zeta) \in \mathcal{P}_2^{n+1}(\mathbb{R}^d)} \quad & F_1(\mu_1) + F_2(\mu_2) + \ldots + F_n(\mu_n) \\ \text{subject to} \quad & \mu_i = \zeta \quad \text{for all } i \in [n] \end{aligned}$$

# Present Work: Distributed Algorithm

$$\underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf} \; F_1(\mu) + F_2(\mu) + \ldots + F_n(\mu)$$

**Main idea:**

$$\Big\Downarrow \text{ re-write}$$

$$\underset{(\mu_1,\ldots,\mu_n,\zeta) \in \mathcal{P}_2^{n+1}(\mathbb{R}^d)}{\arg\inf} \quad F_1(\mu_1) + F_2(\mu_2) + \ldots + F_n(\mu_n)$$

$$\text{subject to} \quad \mu_i = \zeta \quad \text{for all } i \in [n]$$

Define Wasserstein augmented Lagrangian:

$$L_\alpha(\mu_1,\ldots,\mu_n,\zeta,\nu_1,\ldots,\nu_n) := \sum_{i=1}^{n} \left\{ F_i(\mu_i) + \underbrace{\frac{\alpha}{2}}_{\text{regularization} > 0} W^2(\mu_i,\zeta) + \int_{\mathbb{R}^d} \underbrace{\nu_i(\boldsymbol{\theta})}_{\text{Lagrange multipliers}} (\mathrm{d}\mu_i - \mathrm{d}\zeta) \right\}$$

# Proposed Consensus ADMM

$$\mu_i^{k+1} = \underset{\mu_i \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf}\, L_\alpha\big(\mu_1, \ldots, \mu_n, \zeta^k, \nu_1^k, \ldots, \nu_n^k\big)$$

$$\zeta^{k+1} = \underset{\zeta \in \mathcal{P}_2(\mathbb{R}^d)}{\arg\inf}\, L_\alpha\big(\mu_1^{k+1}, \ldots, \mu_n^{k+1}, \zeta, \nu_1^k, \ldots, \nu_n^k\big)$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\big(\mu_i^{k+1} - \zeta^{k+1}\big)$$

where $i \in [n]$, $k \in \mathbb{N}_0$

# Proposed Consensus ADMM

$$\mu_i^{k+1} = \arg\inf_{\mu_i \in \mathcal{P}_2(\mathbb{R}^d)} L_\alpha\left(\mu_1, \ldots, \mu_n, \zeta^k, \nu_1^k, \ldots, \nu_n^k\right)$$

$$\zeta^{k+1} = \arg\inf_{\zeta \in \mathcal{P}_2(\mathbb{R}^d)} L_\alpha\left(\mu_1^{k+1}, \ldots, \mu_n^{k+1}, \zeta, \nu_1^k, \ldots, \nu_n^k\right)$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\left(\mu_i^{k+1} - \zeta^{k+1}\right) \qquad \text{where } i \in [n],\ k \in \mathbb{N}_0$$

Define

$$\nu_{\text{sum}}^k(\boldsymbol{\theta}) := \sum_{i=1}^n \nu_i^k(\boldsymbol{\theta}), \quad k \in \mathbb{N}_0$$

and simplify the recursions to

$$\mu_i^{k+1} = \operatorname{prox}^W_{\frac{1}{\alpha}\left(F_i(\cdot) + \int \nu_i^k \, \mathrm{d}(\cdot)\right)}\left(\zeta^k\right)$$

$$\zeta^{k+1} = \arg\inf_{\zeta \in \mathcal{P}_2(\mathbb{R}^d)} \left\{ \left( \sum_{i=1}^n W^2\left(\mu_i^{k+1}, \zeta\right) \right) - \frac{2}{\alpha} \int_{\mathbb{R}^d} \nu_{\text{sum}}^k(\boldsymbol{\theta}) \mathrm{d}\zeta \right\}$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\left(\mu_i^{k+1} - \zeta^{k+1}\right)$$

# Proposed Consensus ADMM (contd.)

$$\boxed{\mu_i^{k+1} = \operatorname{prox}^{W}_{\frac{1}{\alpha}\left(F_i(\cdot)+\int \nu_i^k \,\mathrm{d}(\cdot)\right)}\left(\zeta^k\right)}$$

$$\zeta^{k+1} = \underset{\zeta\in\mathcal{P}_2\left(\mathbb{R}^d\right)}{\arg\inf}\left\{\left(\sum_{i=1}^{n} W^2\left(\mu_i^{k+1},\zeta\right)\right) - \frac{2}{\alpha}\int_{\mathbb{R}^d}\nu_{\text{sum}}^k(\boldsymbol{\theta})\mathrm{d}\zeta\right\}$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\left(\mu_i^{k+1} - \zeta^{k+1}\right)$$

Split free energy functionals: $\Phi_i(\mu_i) := F_i(\mu_i) + \int_{\mathbb{R}^d}\nu_i^k \,\mathrm{d}\mu_i$

$\therefore$ Distributed Wasserstein prox $\approx$ time updates of $\dfrac{\partial\tilde{\mu}_i}{\partial t} = -\nabla^W\Phi_i(\tilde{\mu}_i)$

# Proposed Consensus ADMM (contd.)

$$\boxed{\mu_i^{k+1} = \mathrm{prox}^{W}_{\frac{1}{\alpha}\left(F_i(\cdot)+\int \nu_i^k \,\mathrm{d}(\cdot)\right)}\left(\zeta^k\right)}$$

$$\zeta^{k+1} = \underset{\zeta\in\mathcal{P}_2(\mathbb{R}^d)}{\arg\inf}\left\{\left(\sum_{i=1}^{n} W^2\left(\mu_i^{k+1},\zeta\right)\right) - \frac{2}{\alpha}\int_{\mathbb{R}^d}\nu_{\mathrm{sum}}^k(\boldsymbol{\theta})\mathrm{d}\zeta\right\}$$

$$\nu_i^{k+1} = \nu_i^k + \alpha\left(\mu_i^{k+1} - \zeta^{k+1}\right)$$

Split free energy functionals: $\Phi_i(\mu_i) := F_i(\mu_i) + \int_{\mathbb{R}^d} \nu_i^k \,\mathrm{d}\mu_i$

$\therefore$ Distributed Wasserstein prox $\approx$ time updates of $\dfrac{\partial\tilde{\mu}_i}{\partial t} = -\nabla^W \Phi_i(\tilde{\mu}_i)$

**Examples:**

| $\Phi_i(\cdot) = F_i(\cdot) + \int \nu_i^k \mathrm{d}(\cdot)$ | PDE | Name |
|---|---|---|
| $\int_{\mathbb{R}^d}\left(V(\boldsymbol{\theta}) + \nu_i^k(\boldsymbol{\theta})\right)\mathrm{d}\mu_i(\boldsymbol{\theta})$ | $\frac{\partial\widetilde{\mu}_i}{\partial t} = \nabla\cdot\left(\widetilde{\mu}_i\left(\nabla V + \nabla\nu_i^k\right)\right)$ | Liouville equation |
| $\int_{\mathbb{R}^d}\left(\nu_i^k(\boldsymbol{\theta}) + \beta^{-1}\log\mu_i(\boldsymbol{\theta})\right)\mathrm{d}\mu_i(\boldsymbol{\theta})$ | $\frac{\partial\widetilde{\mu}_i}{\partial t} = \nabla\cdot\left(\widetilde{\mu}_i\nabla\nu_i^k\right) + \beta^{-1}\Delta\widetilde{\mu}_i$ | Fokker-Planck equation |
| $\int_{\mathbb{R}^d}\nu_i^k(\boldsymbol{\theta})\mathrm{d}\mu_i(\boldsymbol{\theta}) + \int_{\mathbb{R}^{2d}} U(\boldsymbol{\theta},\boldsymbol{\sigma})\mathrm{d}\mu_i(\boldsymbol{\theta})\mathrm{d}\mu_i(\boldsymbol{\sigma})$ | $\frac{\partial\widetilde{\mu}_i}{\partial t} = \nabla\cdot\left(\widetilde{\mu}_i\left(\nabla\nu_i^k + \nabla\left(U\circledast\widetilde{\mu}_i\right)\right)\right)$ | Propagation of chaos equation |
| $\int_{\mathbb{R}^d}\left(\nu_i^k(\boldsymbol{\theta}) + \frac{\beta^{-1}}{m-1}\mathbf{1}^\top\mu_i^m\right)\mathrm{d}\mu_i(\boldsymbol{\theta}), m>1$ | $\frac{\partial\widetilde{\mu}_i}{\partial t} = \nabla\cdot\left(\widetilde{\mu}_i\nabla\nu_i^k\right) + \beta^{-1}\Delta\widetilde{\mu}_i^m$ | Porous medium equation |

# Discrete Version of the Proposed ADMM

$$
\begin{aligned}
\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W}_{\frac{1}{\alpha}\left(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k,\boldsymbol{\mu}_i\rangle\right)}\left(\boldsymbol{\zeta}^k\right) \\
&= \underset{\boldsymbol{\mu}_i\in\Delta^{N-1}}{\arg\inf}\left\{\min_{\boldsymbol{M}\in\Pi_N(\boldsymbol{\mu}_i,\zeta^k)} \frac{1}{2}\langle \boldsymbol{C}, \boldsymbol{M}\rangle + \frac{1}{\alpha}\left(F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\right)\right\} \\
\boldsymbol{\zeta}^{k+1} &= \underset{\zeta\in\Delta^{N-1}}{\arg\inf}\left\{\left(\sum_{i=1}^{n}\min_{\boldsymbol{M}_i\in\Pi_N(\boldsymbol{\mu}_i^{k+1},\zeta)} \frac{1}{2}\langle \boldsymbol{C}, \boldsymbol{M}_i\rangle\right) - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\text{sum}}^k, \boldsymbol{\zeta}\rangle\right\} \\
\boldsymbol{\nu}_i^{k+1} &= \boldsymbol{\nu}_i^k + \alpha\left(\boldsymbol{\mu}_i^{k+1} - \boldsymbol{\zeta}^{k+1}\right)
\end{aligned}
$$

Euclidean distance matrix

where $N$ is the number of samples

# Discrete Version of the Proposed ADMM

$$
\begin{aligned}
\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W}_{\frac{1}{\alpha}(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle)}(\boldsymbol{\zeta}^k) \\
&= \operatorname*{arg\,inf}_{\boldsymbol{\mu}_i \in \Delta^{N-1}} \left\{ \min_{\boldsymbol{M} \in \Pi_N(\boldsymbol{\mu}_i, \zeta^k)} \frac{1}{2}\langle \boldsymbol{C}, \boldsymbol{M}\rangle + \frac{1}{\alpha}\big(F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\big) \right\} \\
\boldsymbol{\zeta}^{k+1} &= \operatorname*{arg\,inf}_{\boldsymbol{\zeta} \in \Delta^{N-1}} \left\{ \left( \sum_{i=1}^{n} \min_{\boldsymbol{M}_i \in \Pi_N(\boldsymbol{\mu}_i^{k+1}, \boldsymbol{\zeta})} \frac{1}{2}\langle \boldsymbol{C}, \boldsymbol{M}_i\rangle \right) - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\mathrm{sum}}^k, \boldsymbol{\zeta}\rangle \right\} \\
\boldsymbol{\nu}_i^{k+1} &= \boldsymbol{\nu}_i^k + \alpha\big(\boldsymbol{\mu}_i^{k+1} - \boldsymbol{\zeta}^{k+1}\big)
\end{aligned}
$$

**With Sinkhorn regularization:**

Discrete Sinkhorn divergence

$$
\begin{aligned}
\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W_\varepsilon}_{\frac{1}{\alpha}(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle)}(\boldsymbol{\zeta}^k) \\
&= \operatorname*{arg\,inf}_{\boldsymbol{\mu}_i \in \Delta^{N-1}} \left\{ \min_{M \in \Pi_N(\boldsymbol{\mu}_i, \zeta^k)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}, \boldsymbol{M} \right\rangle + \frac{1}{\alpha}\big(F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i\rangle\big) \right\} \\
\boldsymbol{\zeta}^{k+1} &= \operatorname*{arg\,inf}_{\boldsymbol{\zeta} \in \Delta^{N-1}} \left\{ \left( \sum_{i=1}^{n} \min_{\boldsymbol{M}_i \in \Pi_N(\boldsymbol{\mu}_i^{k+1}, \boldsymbol{\zeta})} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}_i, \boldsymbol{M}_i \right\rangle \right) - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\mathrm{sum}}^k, \boldsymbol{\zeta}\rangle \right\} \\
\boldsymbol{\nu}_i^{k+1} &= \boldsymbol{\nu}_i^k + \alpha\big(\boldsymbol{\mu}_i^{k+1} - \boldsymbol{\zeta}^{k+1}\big)
\end{aligned}
$$

# Discrete Version of the Proposed ADMM

$$
\begin{aligned}
\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W}_{\frac{1}{\alpha}(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k,\boldsymbol{\mu}_i\rangle)}(\boldsymbol{\zeta}^k) \\
&= \operatorname*{arg\,inf}_{\boldsymbol{\mu}_i\in\Delta^{N-1}} \left\{ \min_{\boldsymbol{M}\in\Pi_N(\boldsymbol{\mu}_i,\zeta^k)} \frac{1}{2}\langle \boldsymbol{C},\boldsymbol{M}\rangle + \frac{1}{\alpha}\big(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k,\boldsymbol{\mu}_i\rangle\big) \right\} \\
\boldsymbol{\zeta}^{k+1} &= \operatorname*{arg\,inf}_{\zeta\in\Delta^{N-1}} \left\{ \left( \sum_{i=1}^{n} \min_{\boldsymbol{M}_i\in\Pi_N(\boldsymbol{\mu}_i^{k+1},\zeta)} \frac{1}{2}\langle \boldsymbol{C},\boldsymbol{M}_i\rangle \right) - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\text{sum}}^k,\boldsymbol{\zeta}\rangle \right\} \\
\boldsymbol{\nu}_i^{k+1} &= \boldsymbol{\nu}_i^k + \alpha\big(\boldsymbol{\mu}_i^{k+1}-\boldsymbol{\zeta}^{k+1}\big)
\end{aligned}
$$

**With Sinkhorn regularization:**

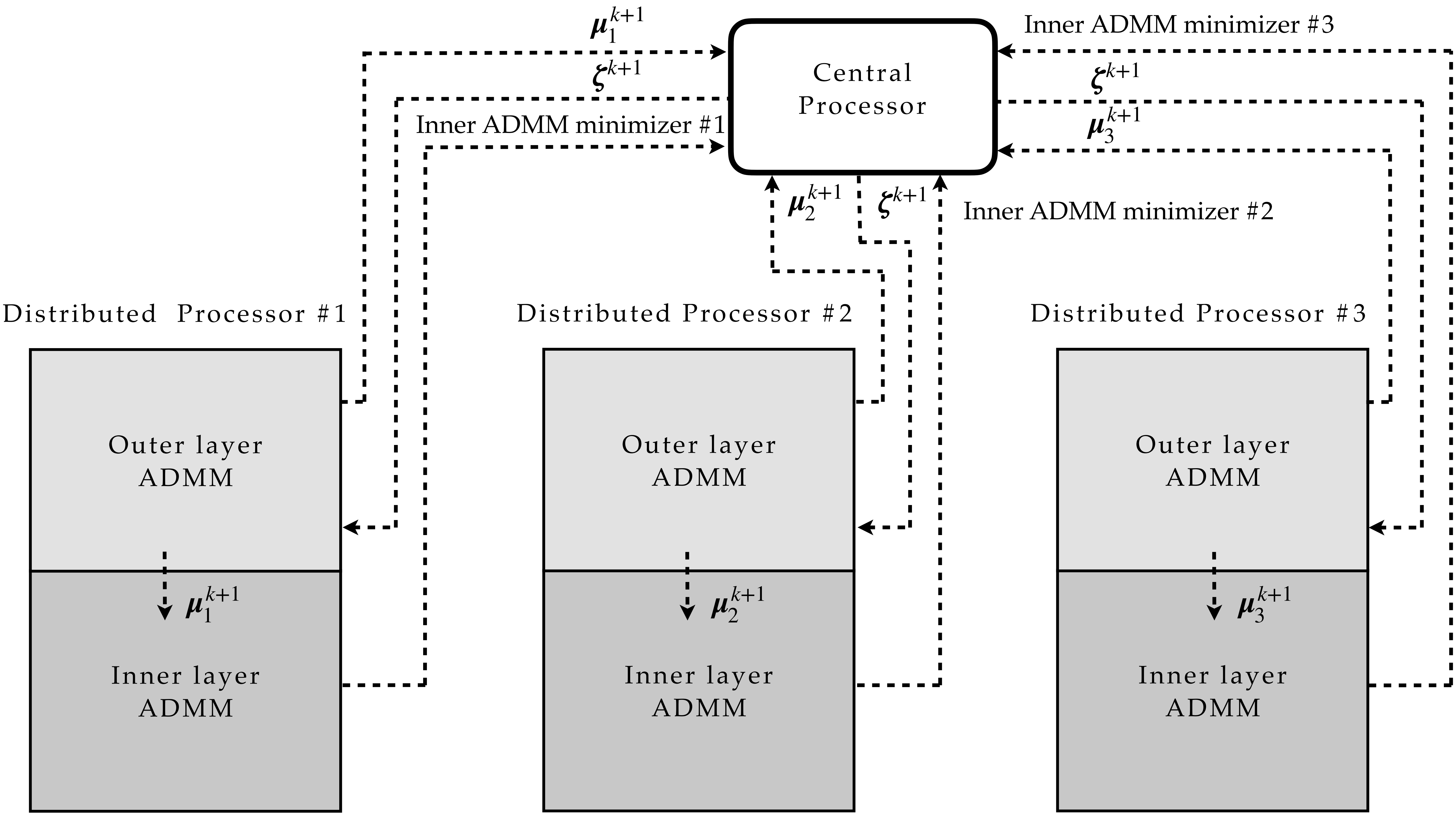
Discrete Sinkhorn divergence

Outer layer ADMM

$$
\begin{aligned}
\boldsymbol{\mu}_i^{k+1} &= \operatorname{prox}^{W_\varepsilon}_{\frac{1}{\alpha}(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k,\boldsymbol{\mu}_i\rangle)}(\boldsymbol{\zeta}^k) \\
&= \operatorname*{arg\,inf}_{\boldsymbol{\mu}_i\in\Delta^{N-1}} \left\{ \min_{M\in\Pi_N(\boldsymbol{\mu}_i,\zeta^k)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon\log\boldsymbol{M},\boldsymbol{M}\right\rangle + \frac{1}{\alpha}\big(F_i(\boldsymbol{\mu}_i)+\langle \boldsymbol{\nu}_i^k,\boldsymbol{\mu}_i\rangle\big) \right\} \\
\boldsymbol{\zeta}^{k+1} &= \operatorname*{arg\,inf}_{\zeta\in\Delta^{N-1}} \left\{ \left( \sum_{i=1}^{n} \min_{\boldsymbol{M}_i\in\Pi_N(\boldsymbol{\mu}_i^{k+1},\zeta)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon\log\boldsymbol{M}_i,\boldsymbol{M}_i\right\rangle \right) - \frac{2}{\alpha}\langle \boldsymbol{\nu}_{\text{sum}}^k,\boldsymbol{\zeta}\rangle \right\} \\
\boldsymbol{\nu}_i^{k+1} &= \boldsymbol{\nu}_i^k + \alpha\big(\boldsymbol{\mu}_i^{k+1}-\boldsymbol{\zeta}^{k+1}\big)
\end{aligned}
$$

Inner layer ADMM

# Overall Schematic



# $\mu_i$ update $\rightsquigarrow$ Outer Consensus (Sinkhorn) ADMM

**Example.** $\Phi(\boldsymbol{\mu}) := \langle \boldsymbol{a}, \boldsymbol{\mu} \rangle$ , $\boldsymbol{a} \in \mathbb{R}^N \backslash \{\boldsymbol{0}\}$, $\boldsymbol{\mu}, \boldsymbol{\zeta} \in \Delta^{N-1}$, $\boldsymbol{\Gamma} := \exp\left(-\boldsymbol{C}/2\varepsilon\right)$, $\varepsilon > 0$

$$\operatorname{prox}_{\frac{1}{\alpha}\Phi}^{W_\varepsilon}(\boldsymbol{\zeta}) = \exp\left(-\frac{1}{\alpha\varepsilon}\boldsymbol{a}\right) \odot \left( \boldsymbol{\Gamma}^\top \left( \boldsymbol{\zeta} \oslash \left( \boldsymbol{\Gamma} \exp\left(-\frac{1}{\alpha\varepsilon}\boldsymbol{a}\right)\right)\right)\right)$$

# $\mu_i$ update ⇝ Outer Consensus (Sinkhorn) ADMM

**Example.** $\Phi(\boldsymbol{\mu}) := \langle \boldsymbol{a}, \boldsymbol{\mu} \rangle$ , $\boldsymbol{a} \in \mathbb{R}^N \backslash \{\mathbf{0}\}$, $\boldsymbol{\mu}, \boldsymbol{\zeta} \in \Delta^{N-1}$, $\boldsymbol{\Gamma} := \exp\left(-\boldsymbol{C}/2\varepsilon\right)$, $\varepsilon > 0$

$$\mathrm{prox}^{W_\varepsilon}_{\frac{1}{\alpha}\Phi}(\boldsymbol{\zeta}) = \exp\left(-\frac{1}{\alpha\varepsilon}\boldsymbol{a}\right) \odot \left(\boldsymbol{\Gamma}^\top\left(\boldsymbol{\zeta} \oslash \left(\boldsymbol{\Gamma}\exp\left(-\frac{1}{\alpha\varepsilon}\boldsymbol{a}\right)\right)\right)\right)$$

**Example.** $G_i(\boldsymbol{\mu_i}) := F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i \rangle$, $\boldsymbol{\zeta}^k \in \Delta^{N-1}$, $k \in \mathbb{N}_0$.

Convex

$$\boldsymbol{\mu}_i^{k+1} = \mathrm{prox}^{W_\varepsilon}_{\frac{1}{\alpha}(F_i(\boldsymbol{\mu}_i) + \langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i \rangle)}\left(\boldsymbol{\zeta}^k\right) = \exp\left(\frac{\boldsymbol{\lambda}_{1i}^{\mathrm{opt}}}{\alpha\varepsilon}\right) \odot \left(\exp\left(-\frac{\boldsymbol{C}^\top}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{0i}^{\mathrm{opt}}}{\alpha\varepsilon}\right)\right)$$

where $\boldsymbol{\lambda}_{0i}^{\mathrm{opt}}, \boldsymbol{\lambda}_{1i}^{\mathrm{opt}} \in \mathbb{R}^N$ solve

$$\exp\left(\frac{\boldsymbol{\lambda}_{0i}^{\mathrm{opt}}}{\alpha\varepsilon}\right) \odot \left(\exp\left(-\frac{\boldsymbol{C}}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{1i}^{\mathrm{opt}}}{\alpha\varepsilon}\right)\right) = \boldsymbol{\zeta}_k,$$

$$\mathbf{0} \in \partial_{\boldsymbol{\lambda}_{1i}^{\mathrm{opt}}} G_i^*\left(-\boldsymbol{\lambda}_{1i}^{\mathrm{opt}}\right) - \exp\left(\frac{\boldsymbol{\lambda}_{1i}^{\mathrm{opt}}}{\alpha\varepsilon}\right) \odot \left(\exp\left(-\frac{\boldsymbol{C}^\top}{2\varepsilon}\right)\exp\left(\frac{\boldsymbol{\lambda}_{0i}^{\mathrm{opt}}}{\alpha\varepsilon}\right)\right).$$

# $\zeta$ update $\rightsquigarrow$ Inner (Euclidean) ADMM

**Theorem.**

Consider the convex problem

$$\begin{aligned}\left(\boldsymbol{u}_1^{\mathrm{opt}},\ldots,\boldsymbol{u}_n^{\mathrm{opt}}\right) = \underset{(\boldsymbol{u}_1,\ldots,\boldsymbol{u}_n)\in\mathbb{R}^{nN}}{\arg\min} & \sum_{i=1}^{n} \left\langle \boldsymbol{\mu}_i^{k+1}, \log\left(\boldsymbol{\Gamma}\exp\left(\boldsymbol{u}_i/\varepsilon\right)\right)\right\rangle \\ \text{subject to} & \sum_{i=1}^{n}\boldsymbol{u}_i = \frac{2}{\alpha}\boldsymbol{\nu}_{\mathrm{sum}}^{k}.\end{aligned} \qquad (\heartsuit)$$

Then

$$\zeta^{k+1} = \exp\left(\boldsymbol{u}_i^{\mathrm{opt}}/\varepsilon\right)\odot\left(\boldsymbol{\Gamma}\left(\boldsymbol{\mu}_i^{k+1}\oslash\left(\boldsymbol{\Gamma}\exp\left(\boldsymbol{u}_i^{\mathrm{opt}}/\varepsilon\right)\right)\right)\right) \;\in\; \Delta^{N-1} \quad \forall\, i\in[n].$$

# $\zeta$ update $\rightsquigarrow$ Inner (Euclidean) ADMM

**Theorem.**

Let $\quad f_i(\boldsymbol{u}_i) := \left\langle \boldsymbol{\mu}_i^{k+1}, \log\left(\boldsymbol{\Gamma} \exp\left(\boldsymbol{u}_i/\varepsilon\right)\right)\right\rangle, \quad \boldsymbol{u}_i \in \mathbb{R}^N, \quad$ for all $i \in [n]$,

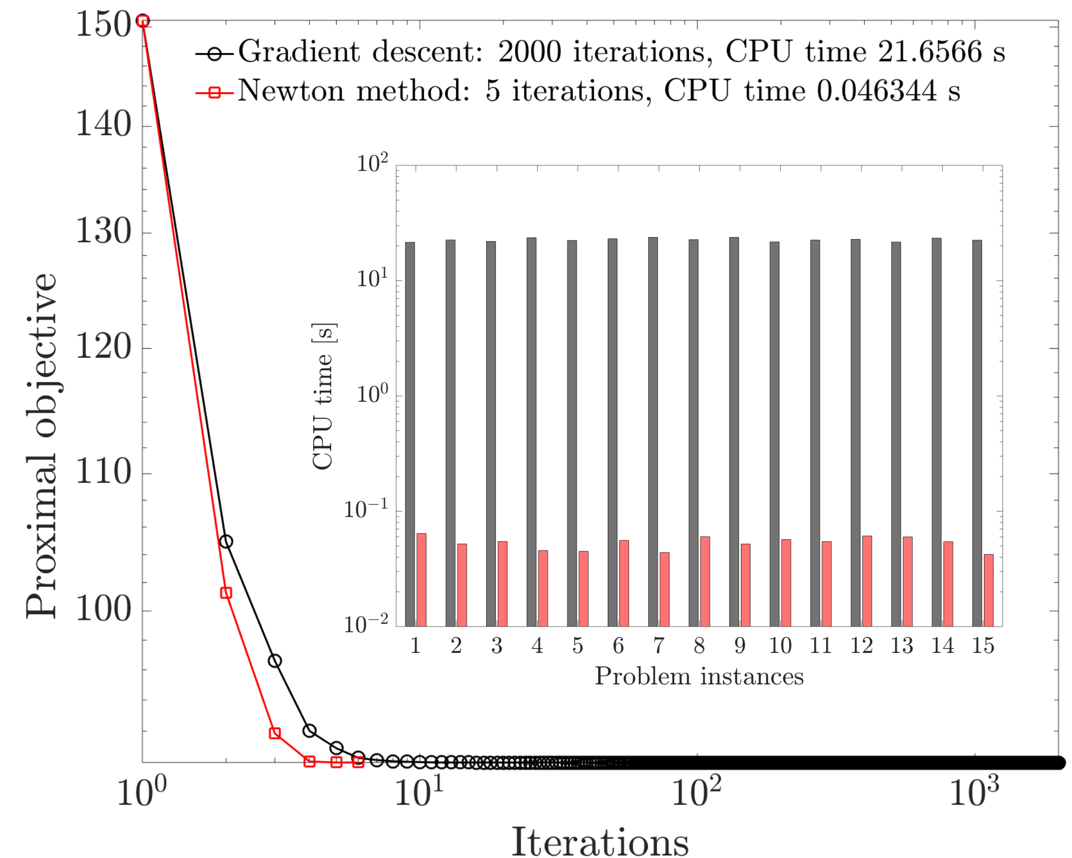
Then the following Euclidean ADMM solves ( ♥ )

$$\boldsymbol{u}_i^{\ell+1} = \boxed{\operatorname{prox}_{\frac{1}{\tau} f_i}^{\|\cdot\|_2}\left(\boldsymbol{z}_i^{\ell} - \widetilde{\boldsymbol{\nu}}_i^{\ell}\right)}$$

No analytical solution, use e.g., Newton's method (has structured Hess)

$$\boldsymbol{z}_i^{\ell+1} = \left(\boldsymbol{u}_i^{\ell+1} - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{u}_i^{\ell+1}\right) + \left(\widetilde{\boldsymbol{\nu}}_i^{\ell} - \frac{1}{n}\sum_{i=1}^{n} \widetilde{\boldsymbol{\nu}}_i^{\ell}\right) + \frac{2}{n\alpha}\boldsymbol{\nu}_{\text{sum}}^{k}$$

$$\widetilde{\boldsymbol{\nu}}_i^{\ell+1} = \widetilde{\boldsymbol{\nu}}_i^{\ell} + \left(\boldsymbol{u}_i^{\ell+1} - \boldsymbol{z}_i^{\ell+1}\right)$$



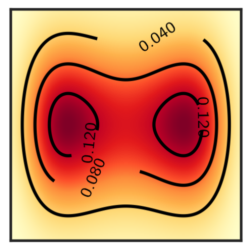
# Experiment #1

Linear Fokker-Planck-Kolmogorov PDE

$$\frac{\partial \mu}{\partial t} = \nabla \cdot (\mu \nabla V) + \beta^{-1} \Delta \mu$$

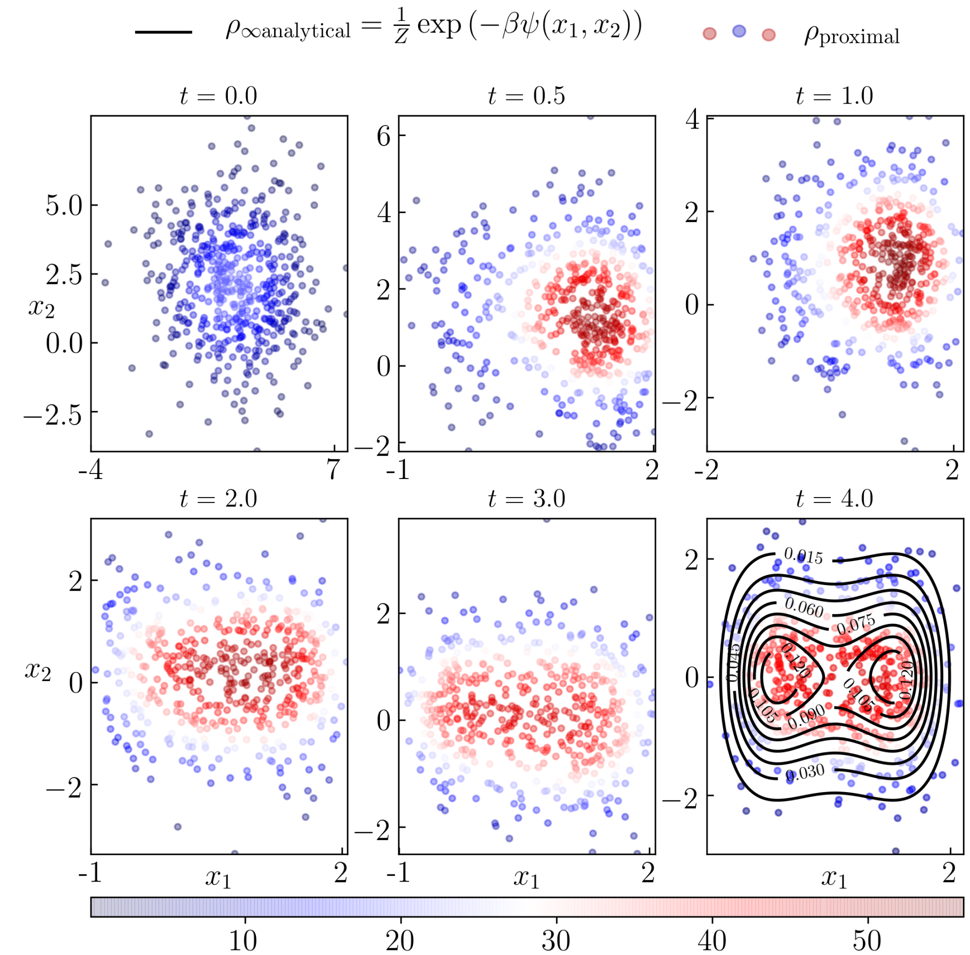
$$V(x_1, x_2) = \frac{1}{4}\left(1 + x_1^4\right) + \frac{1}{2}\left(x_2^2 - x_1^2\right)$$

$$\mu_\infty \propto \exp(-\beta V(x_1, x_2)) \mathrm{d}x_1 \mathrm{d}x_2$$




## Centralized computation:

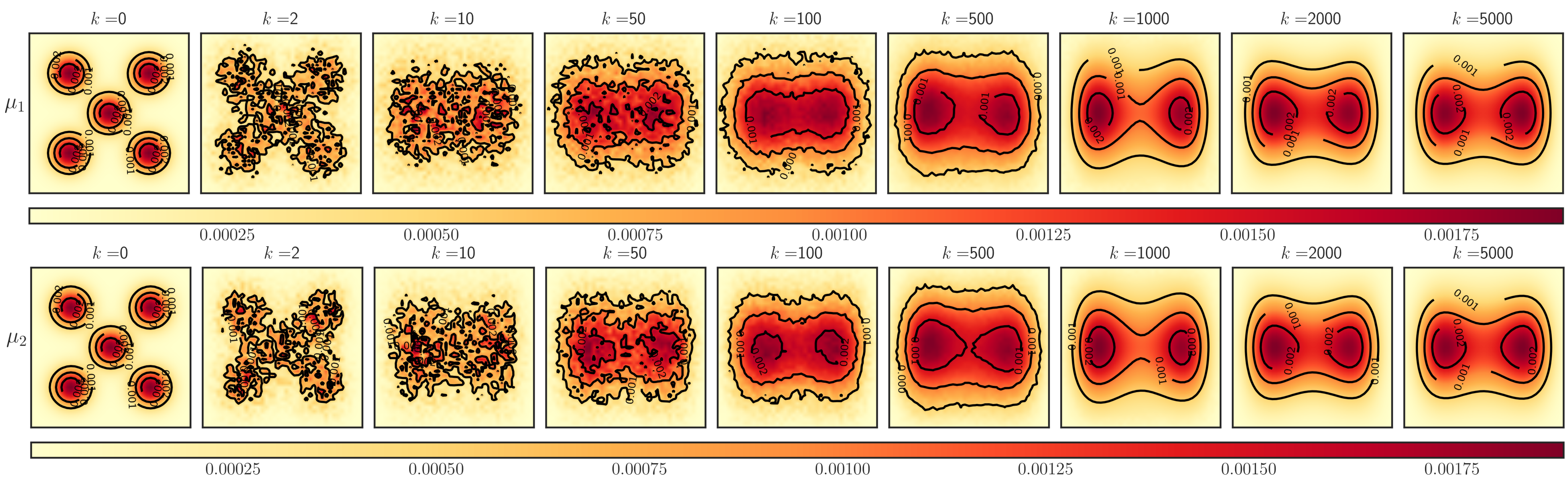
Caluya and Halder, *IEEE Trans. Automatic Control*, 2019




Runtime 99.89 s on Macbook Air 1.1 GHz intel i5 8GB RAM

## Distributed computation:

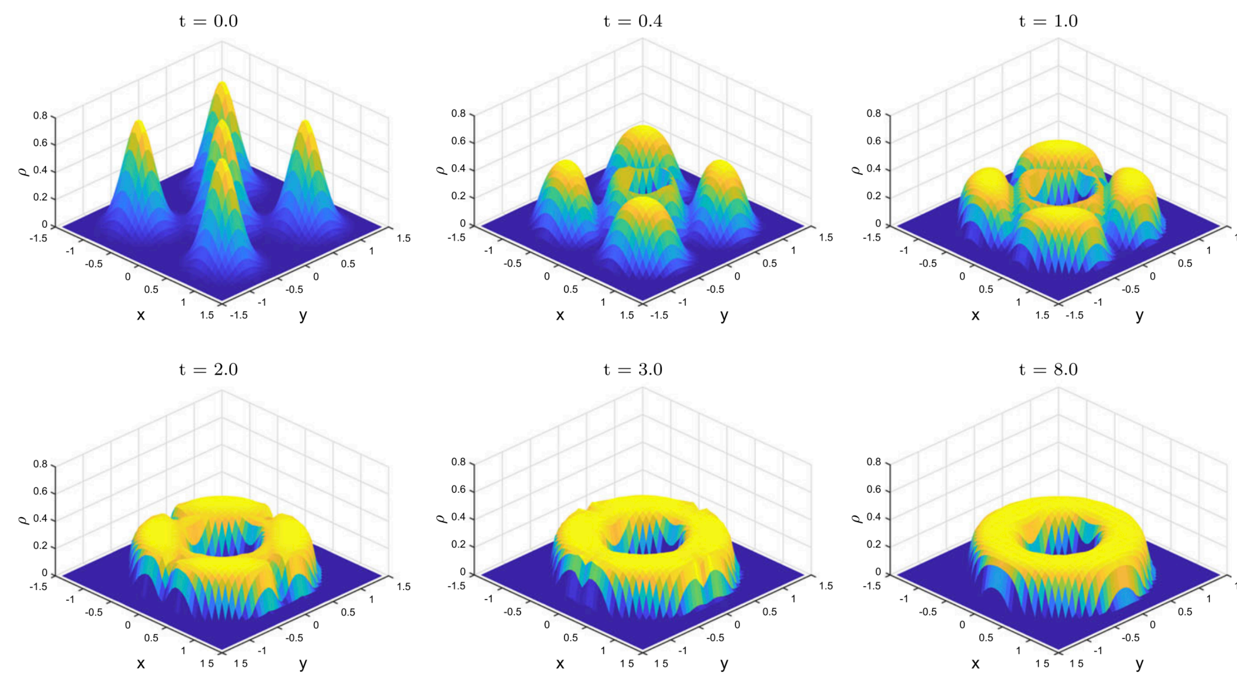
$$F_1(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k, \boldsymbol{\mu} \rangle \qquad F_2(\boldsymbol{\mu}) = \langle \beta^{-1} \log \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$$



# Experiment #2

## Centralized computation:

Carrillo, Craig, Wang and Wei, *FOCM*, 2021




$$\lim_{\beta^{-1}\downarrow 0} \mu_{\infty} = \text{Unif}\,(\mathscr{A})$$

Annulus with inner radius $1/2$ and outer radius $\sqrt{5}/2$

Aggregation-drift-diffusion nonlinear PDE

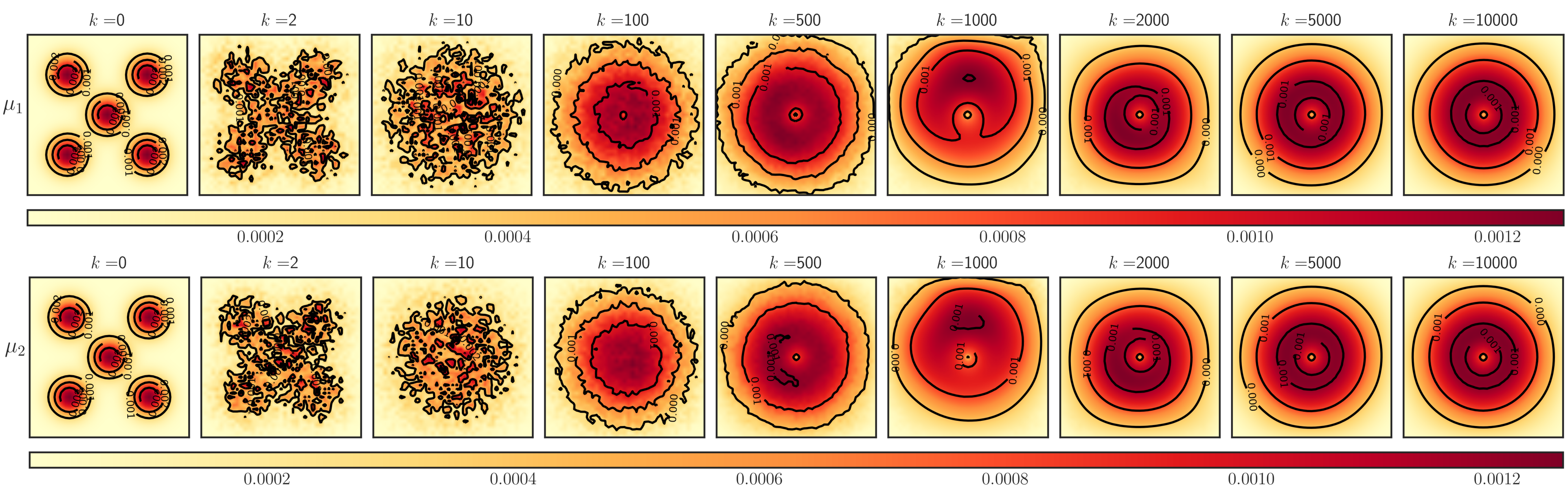
$$\frac{\partial \mu}{\partial t} = \underbrace{\nabla \cdot (\mu \nabla (U * \mu))}_{i=1} + \underbrace{\nabla \cdot (\mu \nabla V) + \beta^{-1} \Delta \mu^2}_{i=2}$$

$$U(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2 - \ln \|\boldsymbol{x}\|_2$$

$$V(\boldsymbol{x}) = -\frac{1}{4} \ln \|\boldsymbol{x}\|_2$$

## Distributed computation:

$$F_1(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}, \boldsymbol{\mu} \rangle \qquad F_2(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k + \beta^{-1} \log \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$$



# Experiment #2 (contd.)

Aggregation-drift-diffusion nonlinear PDE

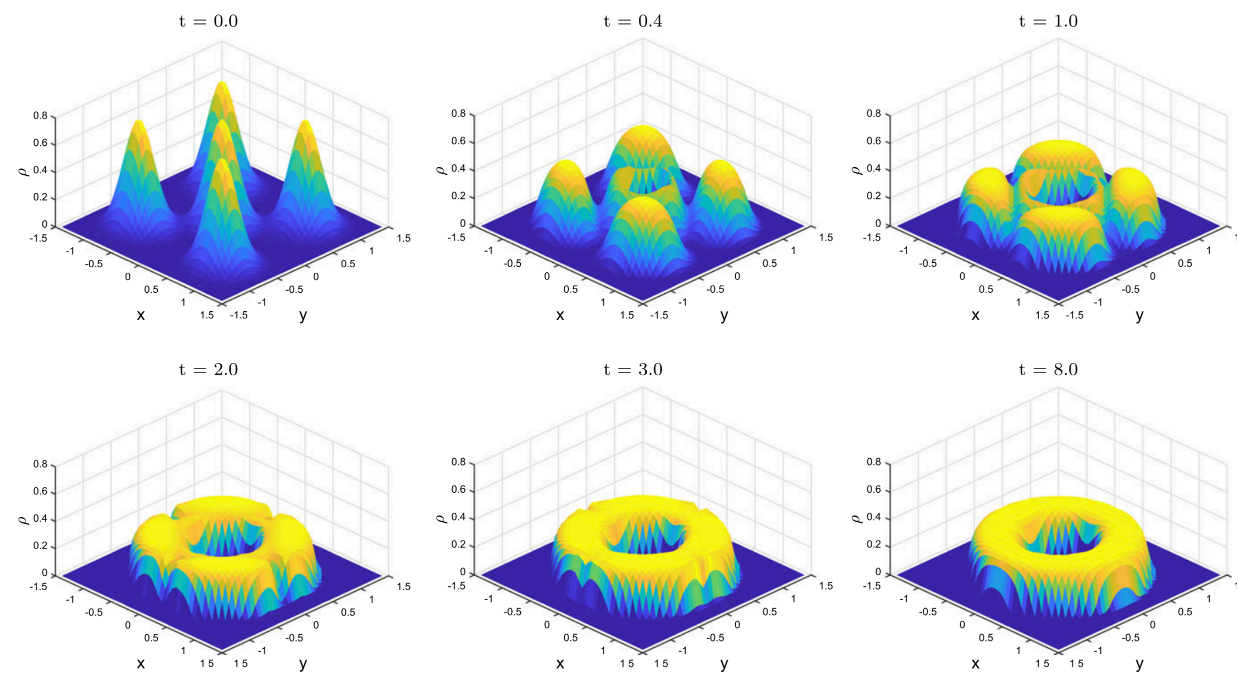
$$\frac{\partial \mu}{\partial t} = \underbrace{\nabla \cdot (\mu \nabla (U * \mu))}_{i=1} + \underbrace{\nabla \cdot (\mu \nabla V) + \beta^{-1} \Delta \mu^2}_{i=2}$$

$$U(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}\|_2^2 - \ln \|\boldsymbol{x}\|_2$$

$$V(\boldsymbol{x}) = -\frac{1}{4} \ln \|\boldsymbol{x}\|_2$$

## Centralized computation:

Carrillo, Craig, Wang and Wei, *FOCM*, 2021

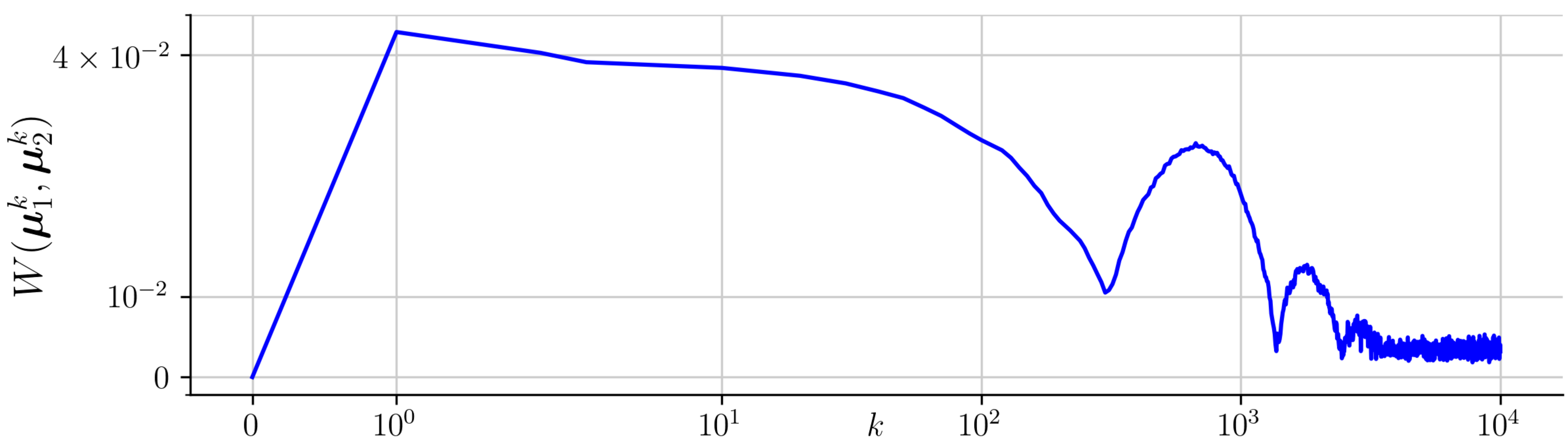



$$\lim_{\beta^{-1} \downarrow 0} \mu_\infty = \mathrm{Unif}\,(\mathscr{A})$$

Annulus with inner radius 1/2 and outer radius $\sqrt{5}/2$

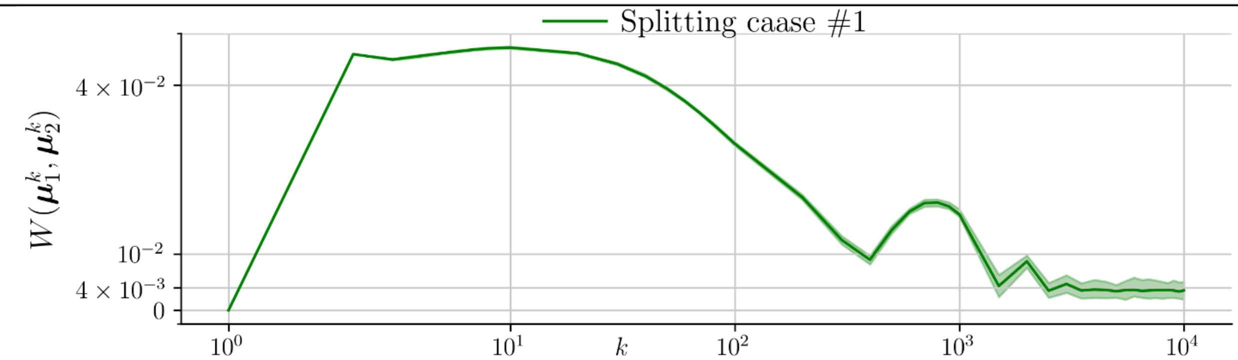
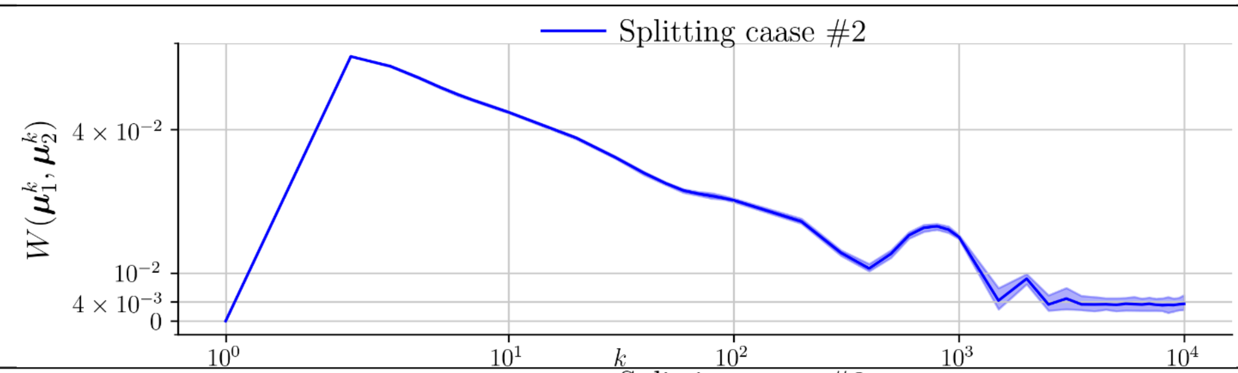
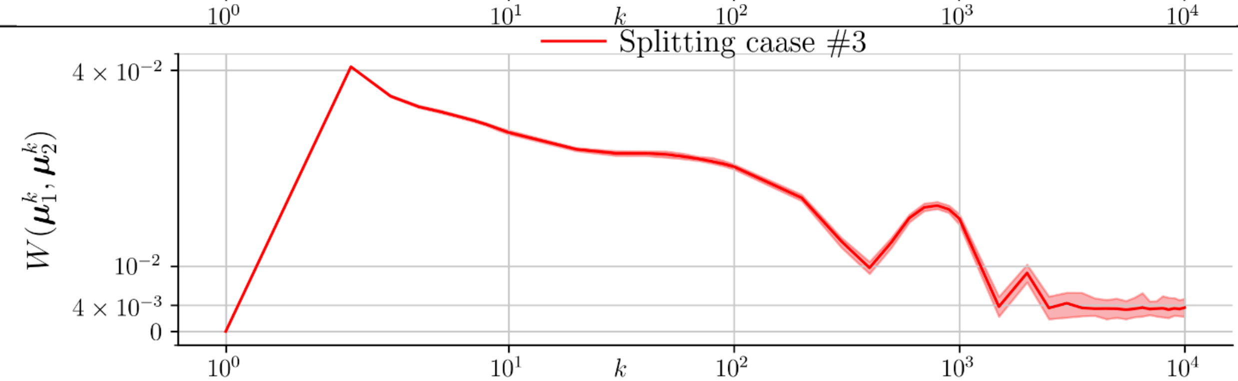
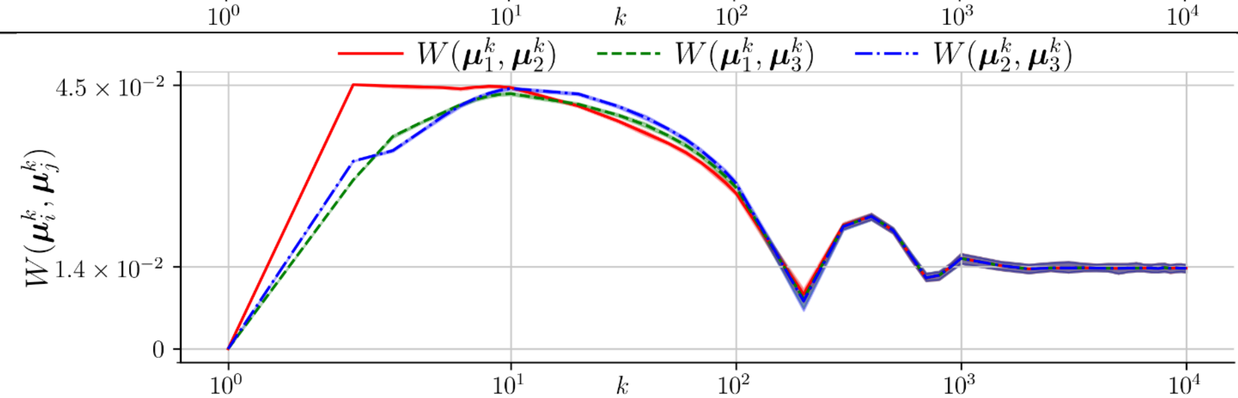
## Distributed computation:

$$F_1(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}, \boldsymbol{\mu} \rangle \qquad F_2(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k + \beta^{-1} \log \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$$



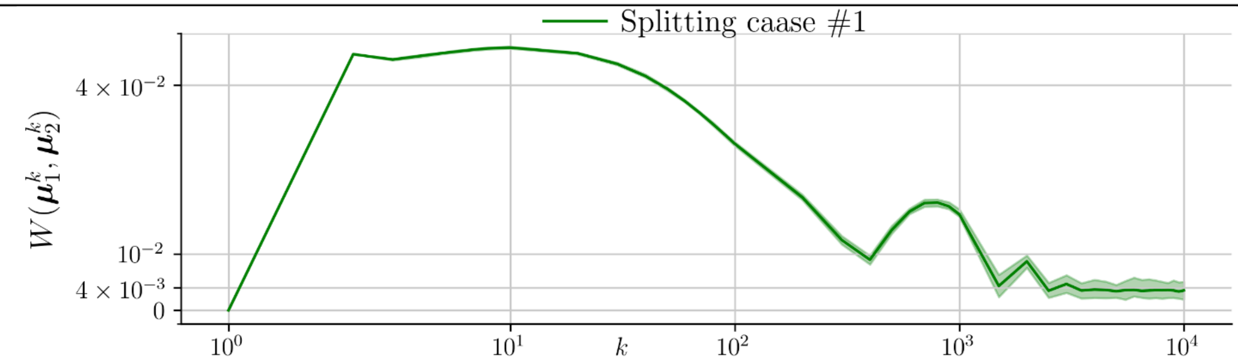
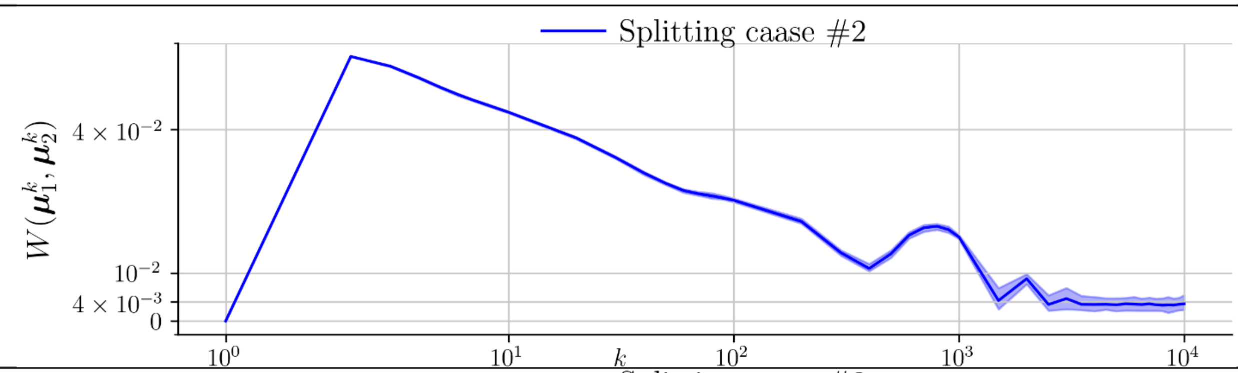
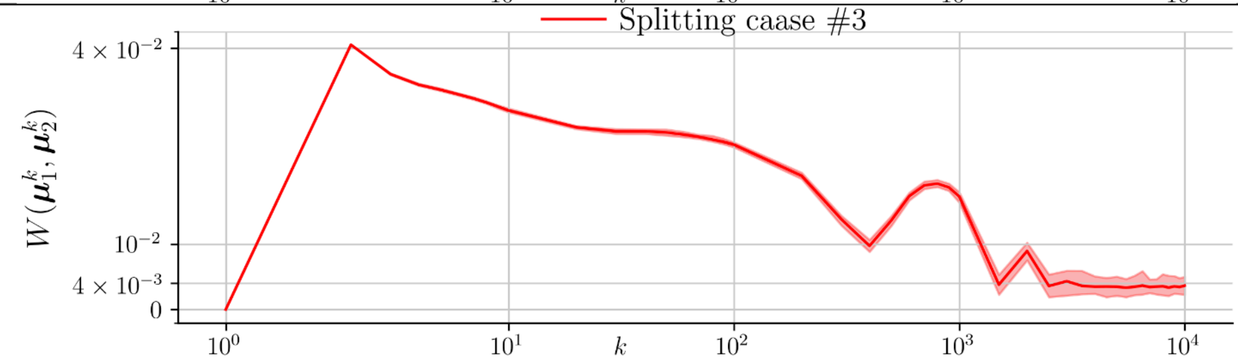
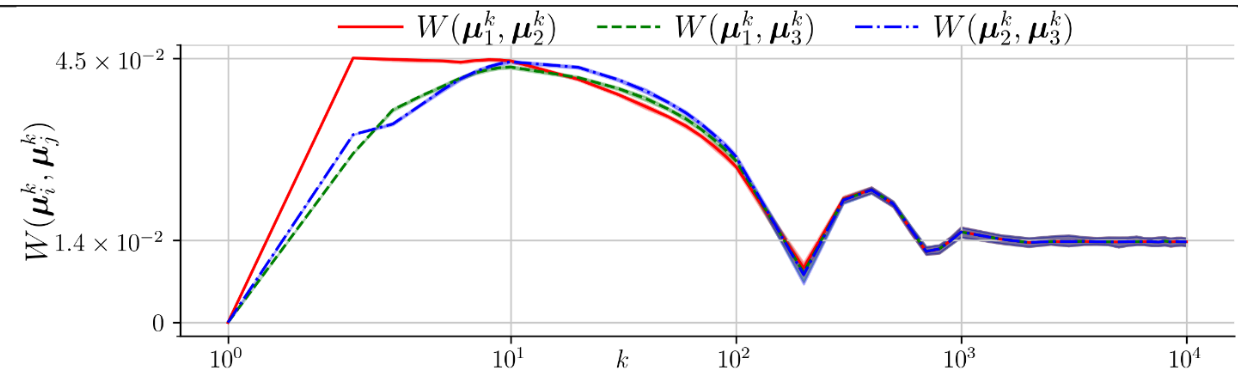
# Experiment # 2 (contd.)

**100 run statistics each of the 4 ways of splitting: ($2^n - n - 1$ ways in general)**

| Splitting case | Functionals | Wasserstein distance |
|---|---|---|
| #1 | $F_1(\boldsymbol{\mu}) = \left\langle \boldsymbol{V}_k + \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \right\rangle$,<br>$F_2(\boldsymbol{\mu}) = \left\langle \boldsymbol{U}_k \boldsymbol{\mu}^k, \boldsymbol{\mu} \right\rangle$ |  |
| #2 | $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}^k + \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \rangle$,<br>$F_2(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k, \boldsymbol{\mu} \rangle$ |  |
| #3 | $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}^k + \boldsymbol{V}_k, \boldsymbol{\mu} \rangle$,<br>$F_2(\boldsymbol{\mu}) = \left\langle \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \right\rangle$ |  |
| #4 | $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k, \boldsymbol{\mu} \rangle$,<br>$F_2(\boldsymbol{\mu}) = \left\langle \boldsymbol{U}_k \boldsymbol{\mu}^k \right\rangle$,<br>$F_3(\boldsymbol{\mu}) = \left\langle \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \right\rangle$ |  |

# Experiment # 2 (contd.)

**100 run statistics each of the 4 ways of splitting: ($2^n - n - 1$ ways in general)**

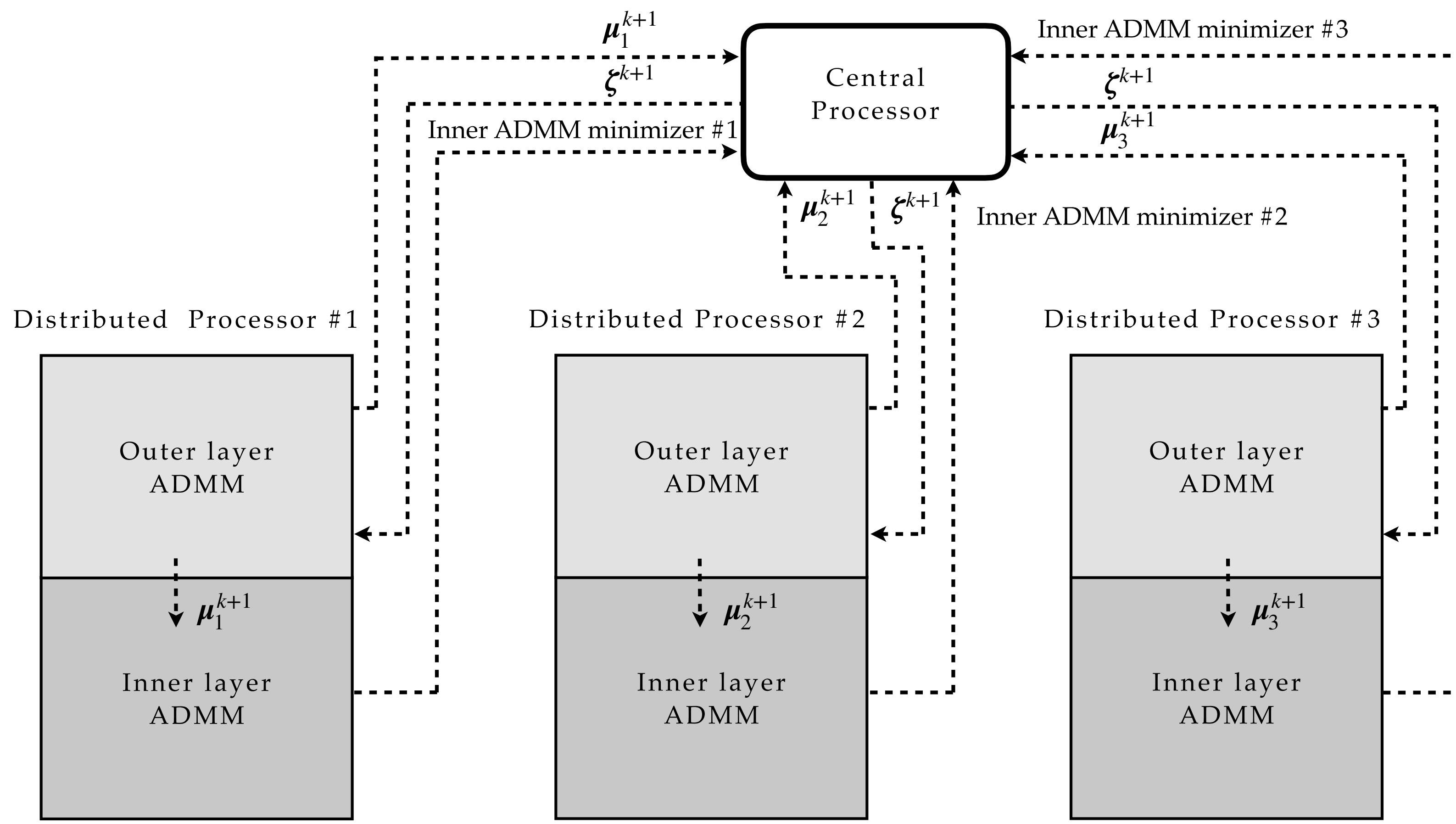
| Splitting case | Functionals | Wasserstein distance |
|---|---|---|
| #1 | $F_1(\boldsymbol{\mu}) = \left\langle \boldsymbol{V}_k + \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \right\rangle,$ <br> $F_2(\boldsymbol{\mu}) = \left\langle \boldsymbol{U}_k \boldsymbol{\mu}^k, \boldsymbol{\mu} \right\rangle$ <br><br> av. runtime = 294.06 s |   |
| #2 | $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}^k + \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \rangle,$ <br> $F_2(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k, \boldsymbol{\mu} \rangle$ <br><br> av. runtime = 285.32 s |   |
| #3 | $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}^k + \boldsymbol{V}_k, \boldsymbol{\mu} \rangle,$ <br> $F_2(\boldsymbol{\mu}) = \left\langle \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \right\rangle$ <br><br> av. runtime = 289.87 s |   |
| #4 | $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k, \boldsymbol{\mu} \rangle,$ <br> $F_2(\boldsymbol{\mu}) = \left\langle \boldsymbol{U}_k \boldsymbol{\mu}^k \right\rangle,$ <br> $F_3(\boldsymbol{\mu}) = \left\langle \beta^{-1}\boldsymbol{\mu}, \boldsymbol{\mu} \right\rangle$ <br><br> av. runtime = 108.99 s |   |

# Summary

Distributed computation for measure-valued optimization

Realizes measure-valued operator splitting

Takes advantage of the existing proximal and JKO type algorithms




# Ongoing

Convergence guarantees for the overall scheme

High dimensional case studies

# Thank You