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

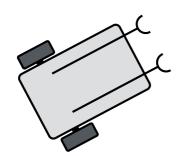
Optimization over the space of measures a.k.a. distributions

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

measure a.k.a. distribution = mass

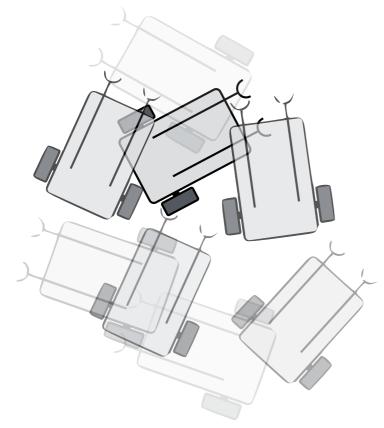
 $mass = density \times volume$

Probability Distribution



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

Probability Distribution

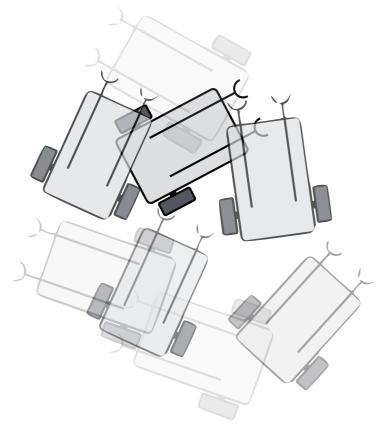


$$\mathbf{x}(t) = \begin{pmatrix} x \\ y \\ \theta \end{pmatrix} \in \mathcal{X} \equiv \mathbb{R}^2 \times \mathbb{S}^1$$

$$\rho\left(x,t\right):\mathcal{X}\times\left[0,\infty\right)\mapsto\mathbb{R}_{\geq0}$$

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

Probability Distribution



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

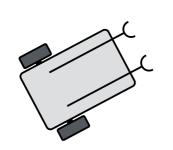
$$\rho(x,t): \mathcal{X} \times [0,\infty) \mapsto \mathbb{R}_{\geq 0}$$

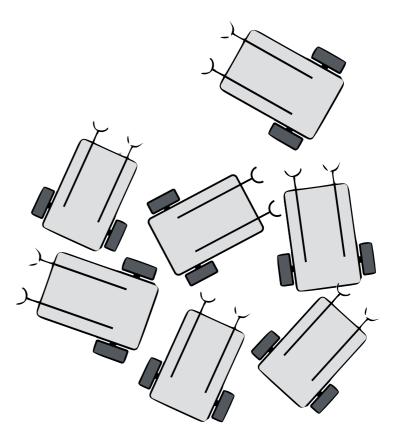
probability measure probability density function

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

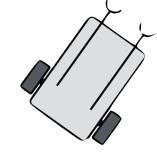
Probability Distribution Population Distribution

Trajectory Generation and Optimal Control





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



population measure population density function

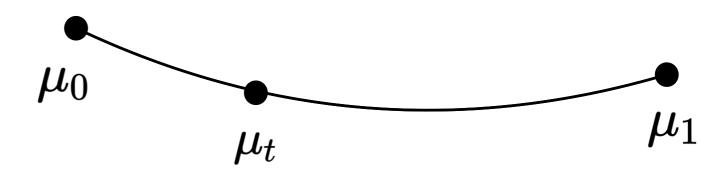
Geometry on the Space of Prob. Measures



2-Wasserstein distance metric

$$egin{align} W_2 := \left(\inf_{\mu,oldsymbol{v}} \left\{rac{1}{2}\int_0^1\int_{\mathcal{X}} \|oldsymbol{v}\|^2 \mathrm{d}\mu \,\mathrm{d}t
ight\}
ight)^{1/2} \ & ext{subject to} \quad rac{\partial \mu}{\partial t} = -
abla \cdot (\muoldsymbol{v}), \; \mu(t=0,\cdot) = \mu_0, \; \mu(t=1,\cdot) = \mu_1 \end{aligned}$$

Geometry on the Space of Prob. Measures



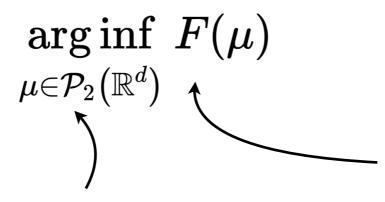
2-Wasserstein distance metric

$$egin{align} W_2 := \left(\inf_{\mu,oldsymbol{v}} \left\{rac{1}{2}\int_0^1\int_{\mathcal{X}}\|oldsymbol{v}\|^2\mathrm{d}\mu\,\mathrm{d}t
ight\}
ight)^{1/2} \ & ext{subject to} \quad rac{\partial\mu}{\partial t} = -
abla\cdot(\muoldsymbol{v}),\; \mu(t=0,\cdot) = \mu_0,\; \mu(t=1,\cdot) = \mu_1 \end{split}$$

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

$$\mu_t = rg\inf_{
u \in \mathcal{P}_2(\mathcal{X})} igg\{ (1-t)W_2^2(\mu_0,
u) + tW_2^2(\mu_1,
u) igg\}$$
 manifold of probability measures supported on \mathcal{X} with finite second moments

Measure-valued Optimization Problems



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

$$rac{\partial \mu}{\partial t} = -
abla_{W_2}^{W_2} F(\mu) :=
abla \cdot \left(\mu
abla rac{\delta F}{\delta \mu}
ight) \qquad (\star)$$

Wasserstein gradient

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

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

Connection with Wasserstein Gradient Flows

$$rac{\partial \mu}{\partial t} = -
abla_{W_2}^{W_2} F(\mu) :=
abla \cdot \left(\mu
abla rac{\delta F}{\delta \mu}
ight) \qquad (\star)$$

Wasserstein gradient

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

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

Gradient Flows

Gradient Flow in \mathcal{X}

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

Recursion:

$$\mathbf{x}_{k} = \mathbf{x}_{k-1} - h\nabla f(\mathbf{x}_{k})$$

$$= \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{arg min}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{x}_{k-1}\|_{2}^{2} + hf(\mathbf{x}) \right\}$$

$$=: \operatorname{prox}_{hf}^{\|\cdot\|_{2}}(\mathbf{x}_{k-1})$$

Convergence:

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

f as Lyapunov function:

$$\frac{\mathrm{d}}{\mathrm{d}t}f = -\parallel \nabla f \parallel_2^2 \le 0$$

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

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

Recursion:

$$= \mathbf{x}_{k-1} - h\nabla f(\mathbf{x}_k)$$

$$= \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{arg min}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{x}_{k-1}\|_{2}^{2} + hf(\mathbf{x}) \right\}$$

$$= : \operatorname{prox}_{hf}^{\|\cdot\|_{2}}(\mathbf{x}_{k-1})$$

$$= : \operatorname{prox}_{hF}^{W}(\mu_{k-1})$$

$$= : \operatorname{prox}_{hF}^{W}(\mu_{k-1})$$

$$= : \operatorname{prox}_{hF}^{W}(\mu_{k-1})$$

Convergence:

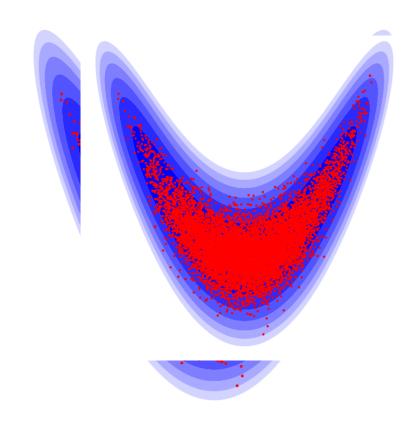
$$\mu_k \rightarrow \mu(\cdot, t = kh)$$
 as $h \downarrow 0$

F as Lyapunov functional:

$$rac{\mathrm{d}}{\mathrm{d}t}F = -\mathbb{E}_{\mu}igg[igg\|
ablarac{\delta F}{\delta\mu}igg\|_2^2igg] \ \le \ 0$$

Motivating Applications

Langevin sampling from an unnormalized prior



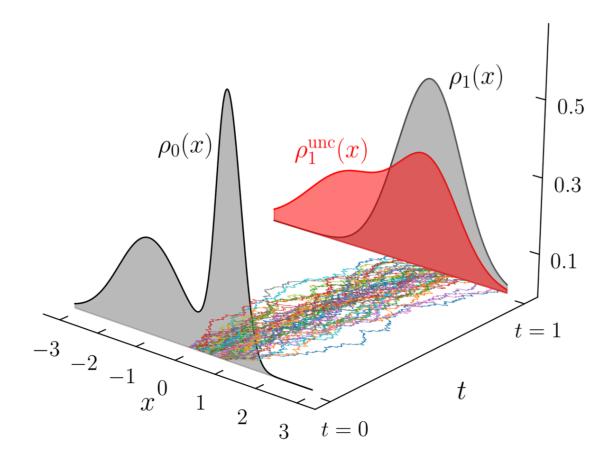
Stramer and Tweedie, *Methodology and Computing* in Applied Probability, 1999

Jarner and Hansen, Stochastic Processes and their Applications, 2000

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

Vempala and Wibisino, NeurIPS, 2019

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



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

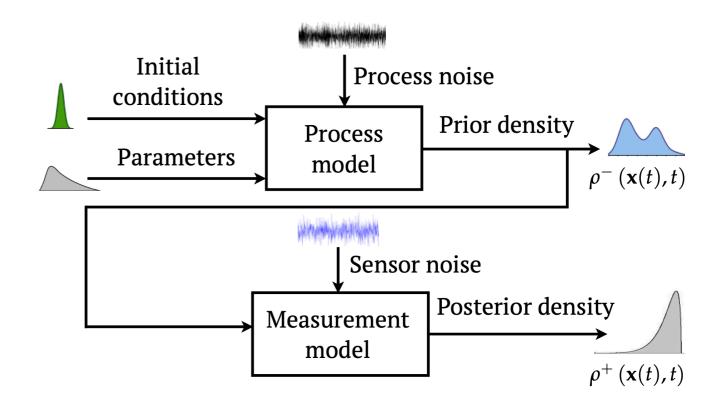
Caluya and Halder, *IEEE Transactions on Automatic Control*, 2021

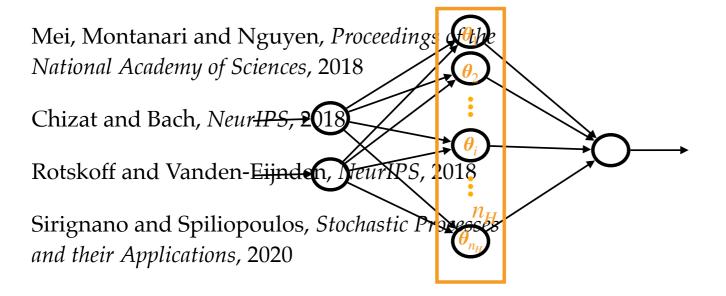
Motivating Applications (contd.)

Mean field learning dynamics in neural networks

 $\begin{array}{c} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_l \\ \vdots \\ \theta_{l} \\ \vdots$

Prediction and estimation of time-varying joint state probability densities





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

For quadratic loss:

$$F(\mu) = F_0 + \int_{\mathbb{R}^p} V(oldsymbol{ heta}) \mathrm{d}\mu(oldsymbol{ heta}) + \int_{\mathbb{R}^{2p}} U(oldsymbol{ heta}, ilde{oldsymbol{ heta}}) \mathrm{d}\mu(oldsymbol{ heta}) \mathrm{d}\mu(oldsymbol{ heta})$$

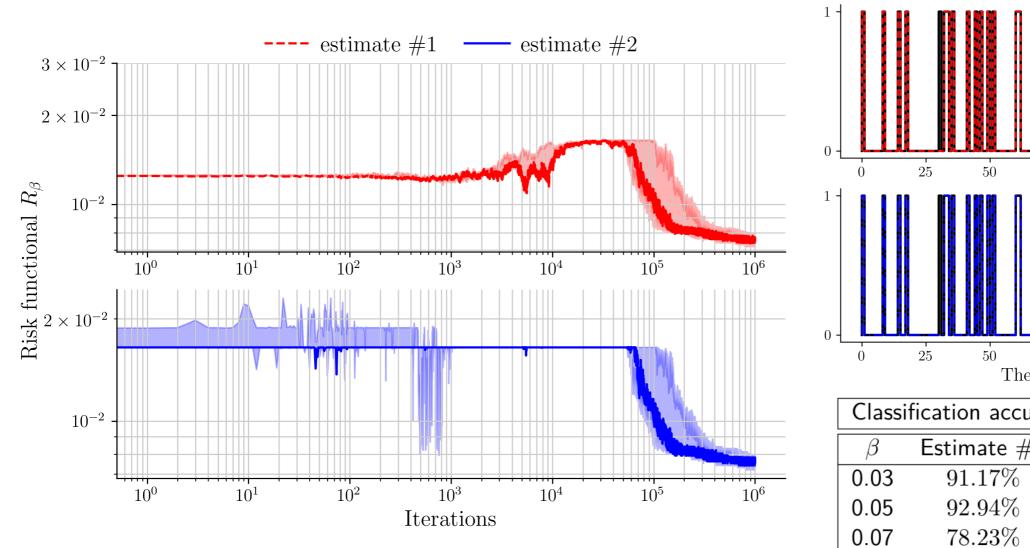
depend on activation functions of the NN

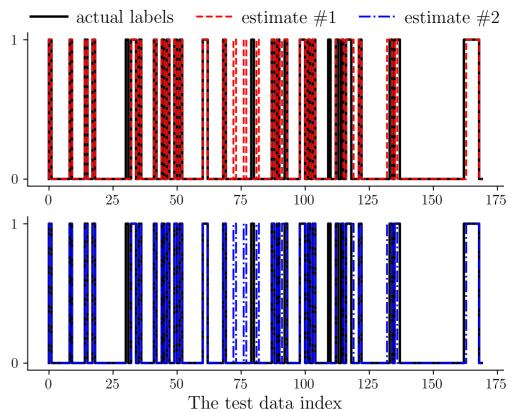
Neuronal population measure dynamics: $\frac{\partial \mu}{\partial t} = \nabla \cdot \left(\mu \nabla \frac{\delta F}{\delta \mu} \right) =: -\nabla^{W_2} F(\mu)$

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

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				
β	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

$$rg \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} F_1(\mu) + F_2(\mu) + \ldots + F_n(\mu)$$

Present Work: Distributed Algorithm

$$rg\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}F_1(\mu)+F_2(\mu)+\ldots+F_n(\mu)$$
 $re ext{-write}$

Main idea:

$$egin{argin} rg \inf \ (\mu_1, \ldots, \mu_n, \zeta) \in \mathcal{P}_2^{n+1}(\mathbb{R}^d) \end{array} F_1(\mu_1) + F_2(\mu_2) + \ldots + F_n(\mu_n) \ ext{subject to} \qquad \mu_i = \zeta \quad ext{for all } i \in [n] \end{cases}$$

Present Work: Distributed Algorithm

$$rg\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)}F_1(\mu)+F_2(\mu)+\ldots+F_n(\mu)$$
 $re ext{-write}$

Main idea:

$$egin{argin} rg \inf \ \mu_1, \ldots, \mu_n, \zeta) \in \mathcal{P}_2^{n+1}(\mathbb{R}^d) \end{array} F_1(\mu_1) + F_2(\mu_2) + \ldots + F_n(\mu_n) \ ext{subject to} \ \mu_i = \zeta \quad ext{for all } i \in [n] \end{aligned}$$

Define Wasserstein augmented Lagrangian:

$$L_{lpha}(\mu_1,\ldots,\mu_n,\zeta,
u_1,\ldots,
u_n) := \sum_{i=1}^n iggl\{ F_i(\mu_i) + rac{lpha}{2} W^2(\mu_i,\zeta) + \int_{\mathbb{R}^d}
u_i(oldsymbol{ heta}) (\mathrm{d}\mu_i - \mathrm{d}\zeta) iggr\}$$
 regularization > 0 Lagrange multipliers

Proposed Consensus ADMM

$$egin{aligned} \mu_i^{k+1} &= rg\inf_{\mu_i \in \mathcal{P}_2(\mathbb{R}^d)} L_lphaig(\mu_1,\ldots,\mu_n,\zeta^k,
u_1^k,\ldots,
u_n^kig) \ \zeta^{k+1} &= rg\inf_{\zeta \in \mathcal{P}_2(\mathbb{R}^d)} L_lphaig(\mu_1^{k+1},\ldots,\mu_n^{k+1},\zeta,
u_1^k,\ldots,
u_n^kig) \ arphi^{k+1} &=
u_i^k + lphaig(\mu_i^{k+1} - \zeta^{k+1}ig) \end{aligned} \qquad ext{where } i \in [n], k \in \mathbb{N}_0$$

Proposed Consensus ADMM

$$egin{aligned} \mu_i^{k+1} &= rg\inf_{\mu_i \in \mathcal{P}_2(\mathbb{R}^d)} L_lphaig(\mu_1,\ldots,\mu_n,\zeta^k,
u_1^k,\ldots,
u_n^kig) \ \zeta^{k+1} &= rg\inf_{\zeta \in \mathcal{P}_2(\mathbb{R}^d)} L_lphaig(\mu_1^{k+1},\ldots,\mu_n^{k+1},\zeta,
u_1^k,\ldots,
u_n^kig) \
u_i^{k+1} &=
u_i^k + lphaig(\mu_i^{k+1} - \zeta^{k+1}ig) \qquad \qquad ext{where } i \in [n], k \in \mathbb{N}_0 \end{aligned}$$

Define

$$u_{ ext{sum}}^k\left(oldsymbol{ heta}
ight) := \sum_{i=1}^n
u_i^k(oldsymbol{ heta}), \quad k \in \mathbb{N}_0$$

and simplify the recursions to

$$egin{aligned} \mu_i^{k+1} &= \operatorname{prox}_{rac{1}{lpha} \left(F_i(\cdot) + \int
u_i^k \, \mathrm{d}(\cdot)
ight)}^W \left(\zeta^k
ight) \ \zeta^{k+1} &= rg \inf_{\zeta \in \mathcal{P}_2\left(\mathbb{R}^d
ight)}^W \left\{ \left(\sum_{i=1}^n W^2ig(\mu_i^{k+1}, \zetaig)
ight) - rac{2}{lpha} \int_{\mathbb{R}^d}
u_{ ext{sum}}^k(oldsymbol{ heta}) \, \mathrm{d}\zeta
ight\} \
u_i^{k+1} &=
u_i^k + lpha ig(\mu_i^{k+1} - \zeta^{k+1}ig) \end{aligned}$$

Proposed Consensus ADMM (contd.)

$$egin{aligned} \left(\mu_i^{k+1} &= \operatorname{prox}_{rac{1}{lpha}\left(F_i(\cdot) + \int
u_i^k \operatorname{d}(\cdot)
ight)}^W \left(\zeta^k
ight) \ \zeta^{k+1} &= rg\inf_{\zeta \in \mathcal{P}_2\left(\mathbb{R}^d
ight)} \left\{ \left(\sum_{i=1}^n W^2ig(\mu_i^{k+1},\zetaig)
ight) - rac{2}{lpha} \int_{\mathbb{R}^d}
u_{ ext{sum}}^k(oldsymbol{ heta}) \operatorname{d}\zeta
ight\} \
u_i^{k+1} &=
u_i^k + lphaig(\mu_i^{k+1} - \zeta^{k+1}ig) \end{aligned}$$

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

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

Proposed Consensus ADMM (contd.)

$$egin{aligned} egin{aligned} egi$$

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

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

Examples:

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

Discrete Version of the Proposed ADMM

$$oldsymbol{\mu}_i^{k+1} = \operatorname{prox}_{rac{1}{lpha}\left(F_i(oldsymbol{\mu}_i) + \left\langle oldsymbol{
u}_i^k, oldsymbol{\mu}_i
ight)}{\left\{ egin{align*}{c} \operatorname{min}_{oldsymbol{M} \in \Pi_N(oldsymbol{\mu}_i, \zeta^k)} rac{1}{2} \left\langle oldsymbol{C}, oldsymbol{M}
ight) + rac{1}{lpha} \left(F_i(oldsymbol{\mu}_i) + \left\langle oldsymbol{
u}_i^k, oldsymbol{\mu}_i
ight)
ight) \\ oldsymbol{\zeta}^{k+1} = rg \inf_{oldsymbol{\zeta} \in \Delta^{N-1}} \left\{ \left(\sum_{i=1}^n \min_{oldsymbol{M}_i \in \Pi_N(oldsymbol{\mu}_i^{k+1}, oldsymbol{\zeta})} rac{1}{2} \left\langle oldsymbol{C}, oldsymbol{M}_i
ight) - rac{2}{lpha} \left\langle oldsymbol{
u}_{\mathrm{sum}}^k, oldsymbol{\zeta}
ight) \\ oldsymbol{
u}_i^{k+1} = oldsymbol{
u}_i^k + lpha \left(oldsymbol{\mu}_i^{k+1} - oldsymbol{\zeta}^{k+1}
ight) \qquad \qquad \text{where N is the number of samples} \end{cases}$$

Discrete Version of the Proposed ADMM

$$egin{aligned} oldsymbol{\mu}_i^{k+1} &= \operatorname{prox}_{rac{1}{lpha}ig(F_i(oldsymbol{\mu}_i) + ig\langle oldsymbol{
u}_i^k, oldsymbol{\mu}_i ig)}^{W} ig(oldsymbol{\zeta}^k) \ &= rg\inf_{oldsymbol{\mu}_i \in \Delta^{N-1}} igg\{ \min_{oldsymbol{M} \in \Pi_N(oldsymbol{\mu}_i, oldsymbol{\zeta}^k)} rac{1}{2} ig\langle oldsymbol{C}, oldsymbol{M} ig
angle + rac{1}{lpha} ig(F_i(oldsymbol{\mu}_i) + ig\langle oldsymbol{
u}_i^k, oldsymbol{\mu}_i ig
angle ig) igg\} \ oldsymbol{\zeta}^{k+1} &= rg\inf_{oldsymbol{\zeta} \in \Delta^{N-1}} igg\{ igg(\sum_{i=1}^n \min_{oldsymbol{M}_i \in \Pi_N(oldsymbol{\mu}_i^{k+1}, oldsymbol{\zeta})} rac{1}{2} ig\langle oldsymbol{C}, oldsymbol{M}_i ig
angle igg) - rac{2}{lpha} ig\langle oldsymbol{
u}_{ ext{sum}}^k, oldsymbol{\zeta} igg
angle igg\} \ oldsymbol{
u}_i^{k+1} &= oldsymbol{
u}_i^k + lpha ig(oldsymbol{\mu}_i^{k+1} - oldsymbol{\zeta}^{k+1} ig) \ \end{pmatrix}$$

With Sinkhorn regularization:

Discrete Sinkhorn divergence
$$\boldsymbol{\mu}_{i}^{k+1} = \operatorname{prox}_{\frac{1}{\alpha}\left(F_{i}(\boldsymbol{\mu}_{i}) + \left\langle\boldsymbol{\nu}_{i}^{k}, \boldsymbol{\mu}_{i}\right\rangle\right)}^{\boldsymbol{W}_{\varepsilon}} \left(\boldsymbol{\zeta}^{k}\right)$$

$$= \operatorname*{arg inf}_{\boldsymbol{\mu}_{i} \in \Delta^{N-1}} \left\{ \operatorname*{min}_{\boldsymbol{M} \in \Pi_{N}\left(\boldsymbol{\mu}_{i}, \boldsymbol{\zeta}^{k}\right)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}, \boldsymbol{M} \right\rangle + \frac{1}{\alpha} \left(F_{i}(\boldsymbol{\mu}_{i}) + \left\langle\boldsymbol{\nu}_{i}^{k}, \boldsymbol{\mu}_{i}\right\rangle\right) \right\}$$

$$\boldsymbol{\zeta}^{k+1} = \operatorname*{arg inf}_{\boldsymbol{\zeta} \in \Delta^{N-1}} \left\{ \left(\sum_{i=1}^{n} \operatorname*{min}_{\boldsymbol{M}_{i} \in \Pi_{N}\left(\boldsymbol{\mu}_{i}^{k+1}, \boldsymbol{\zeta}\right)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}_{i}, \boldsymbol{M}_{i} \right\rangle \right) - \frac{2}{\alpha} \left\langle\boldsymbol{\nu}_{\operatorname{sum}}^{k}, \boldsymbol{\zeta}\right\rangle \right\}$$

$$\boldsymbol{\nu}_{i}^{k+1} = \boldsymbol{\nu}_{i}^{k} + \alpha \left(\boldsymbol{\mu}_{i}^{k+1} - \boldsymbol{\zeta}^{k+1}\right)$$

Discrete Version of the Proposed ADMM

$$egin{aligned} oldsymbol{\mu}_i^{k+1} &= \operatorname{prox}_{rac{1}{lpha}ig(F_i(oldsymbol{\mu}_i) + ig\langle oldsymbol{
u}_i^k, oldsymbol{\mu}_iig)}^{W} ig(oldsymbol{\zeta}^k) \ &= rg\inf_{oldsymbol{\mu}_i \in \Delta^{N-1}} igg\{ \min_{oldsymbol{M} \in \Pi_N(oldsymbol{\mu}_i, oldsymbol{\zeta}^k)} rac{1}{2} ig\langle oldsymbol{C}, oldsymbol{M}ig
angle + rac{1}{lpha} ig(F_i(oldsymbol{\mu}_i) + ig\langle oldsymbol{
u}_i^k, oldsymbol{\mu}_iig
angle ig) igg\} \ oldsymbol{\zeta}^{k+1} &= rg\inf_{oldsymbol{\zeta} \in \Delta^{N-1}} igg\{ igg(\sum_{i=1}^n \min_{oldsymbol{M}_i \in \Pi_N(oldsymbol{\mu}_i^{k+1}, oldsymbol{\zeta})} rac{1}{2} ig\langle oldsymbol{C}, oldsymbol{M}_i igr
angle igg) - rac{2}{lpha} ig\langle oldsymbol{
u}_{ ext{sum}}^k, oldsymbol{\zeta} igr
angle igg\} \ oldsymbol{
u}_i^{k+1} &= oldsymbol{
u}_i^k + lpha ig(oldsymbol{\mu}_i^{k+1} - oldsymbol{\zeta}^{k+1} ig) \ \end{pmatrix}$$

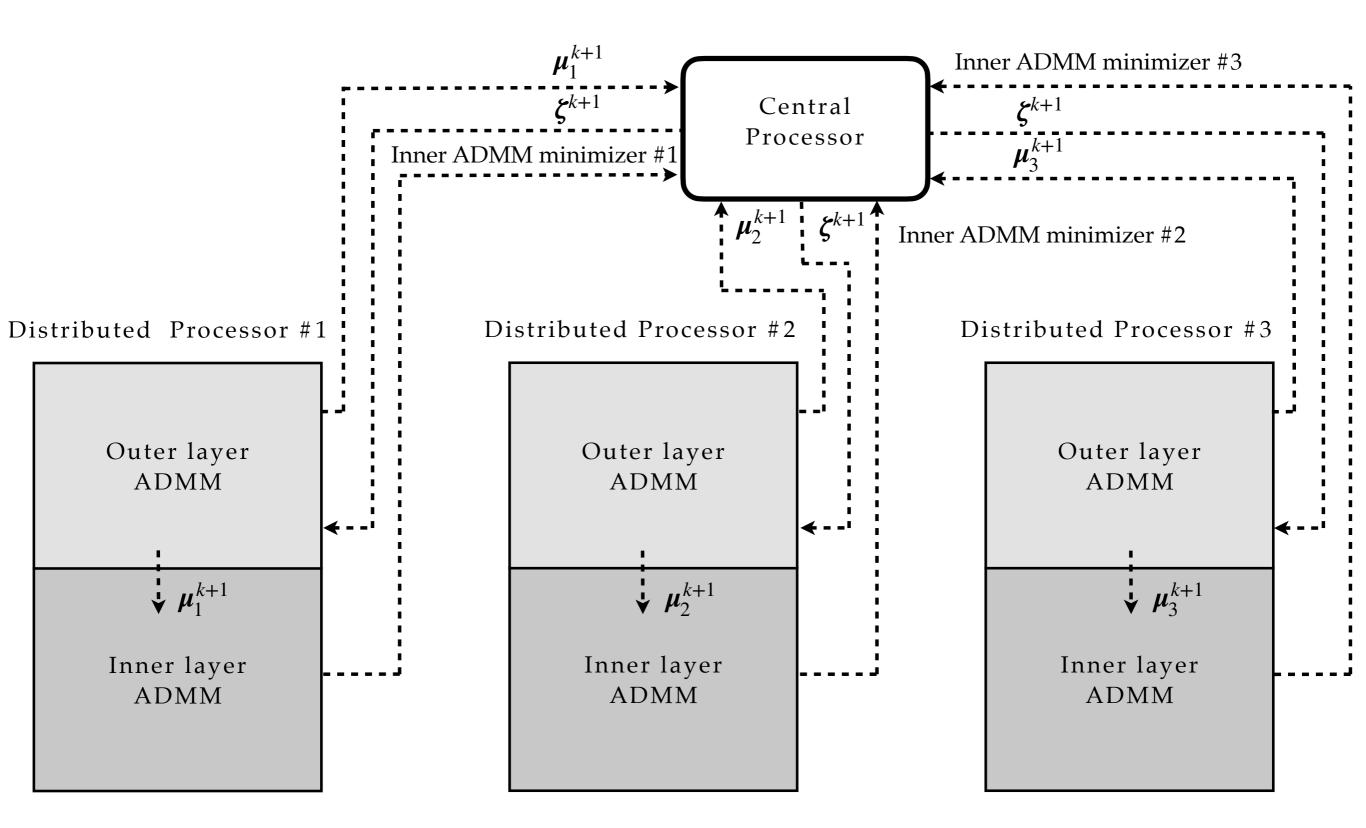
With Sinkhorn regularization:

Discrete Sinkhorn divergence

Outer layer ADMM
$$\boldsymbol{\zeta}^{k+1} = \operatorname*{arg\,inf}_{\boldsymbol{\zeta} \in \Delta^{N-1}} \left\{ \frac{\min\limits_{\boldsymbol{M} \in \Pi_N(\boldsymbol{\mu}_i,\boldsymbol{\zeta}^k)} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}, \boldsymbol{M} \right\rangle + \frac{1}{\alpha} \left(F_i(\boldsymbol{\mu}_i) + \left\langle \boldsymbol{\nu}_i^k, \boldsymbol{\mu}_i \right\rangle \right) \right\} \\ \boldsymbol{\zeta}^{k+1} = \operatorname*{arg\,inf}_{\boldsymbol{\zeta} \in \Delta^{N-1}} \left\{ \left(\sum_{i=1}^n \min\limits_{\boldsymbol{M}_i \in \Pi_N(\boldsymbol{\mu}_i^{k+1},\boldsymbol{\zeta})} \left\langle \frac{1}{2}\boldsymbol{C} + \varepsilon \log \boldsymbol{M}_i, \boldsymbol{M}_i \right\rangle \right) - \frac{2}{\alpha} \left\langle \boldsymbol{\nu}_{\mathrm{sum}}^k, \boldsymbol{\zeta} \right\rangle \right\} \\ \boldsymbol{\nu}_i^{k+1} = \boldsymbol{\nu}_i^k + \alpha \left(\boldsymbol{\mu}_i^{k+1} - \boldsymbol{\zeta}^{k+1} \right)$$
 Inner layer ADMN

layer **ADMM**

Overall Schematic



μ_i update \rightsquigarrow Outer Consensus (Sinkhorn) ADMM

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

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

μ_i update \rightsquigarrow Outer Consensus (Sinkhorn) ADMM

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

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

Example.
$$G_i(\mu_i):=F_i(\mu_i)+\langle \nu_i^k,\mu_i\rangle,~\zeta^k\in\Delta^{N-1},~k\in\mathbb{N}_0.$$
Convex

$$m{\mu}_i^{k+1} = \mathrm{prox}_{rac{1}{lpha}\left(F_i(m{\mu}_i) + \left\langle m{
u}_i^k, m{\mu}_i
ight
angle} \left(m{\zeta}^k
ight) = \exp\left(rac{m{\lambda}_{1i}^{\mathrm{opt}}}{lpha arepsilon}
ight) \odot \left(\exp\left(-rac{m{C}^ op}{2arepsilon}
ight) \exp\left(rac{m{\lambda}_{0i}^{\mathrm{opt}}}{lpha arepsilon}
ight)
ight)$$

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

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

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

ζupdate → Inner (Euclidean) ADMM

Theorem.

Consider the convex problem

$$(\boldsymbol{u}_{1}^{\text{opt}}, \dots, \boldsymbol{u}_{n}^{\text{opt}}) = \underset{(\boldsymbol{u}_{1}, \dots, \boldsymbol{u}_{n}) \in \mathbb{R}^{nN}}{\operatorname{arg \, min}} \sum_{i=1}^{n} \langle \boldsymbol{\mu}_{i}^{k+1}, \log \left(\boldsymbol{\Gamma} \exp \left(\boldsymbol{u}_{i} / \varepsilon \right) \right) \rangle$$
subject to
$$\sum_{i=1}^{n} \boldsymbol{u}_{i} = \frac{2}{\alpha} \boldsymbol{\nu}_{\text{sum}}^{k}.$$
 (\mathbf{\psi})

Then

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

ζupdate → Inner (Euclidean) ADMM

Theorem.

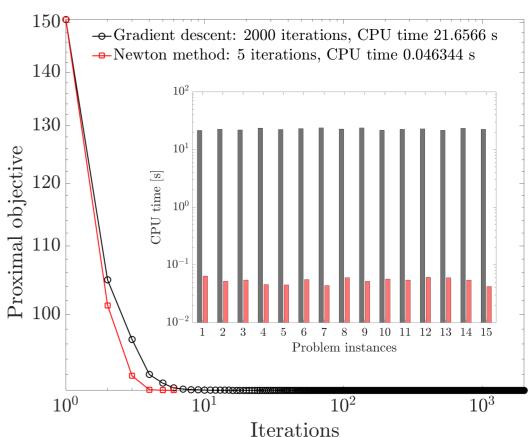
Let
$$f_i(\boldsymbol{u}_i) := \langle \boldsymbol{\mu}_i^{k+1}, \log \left(\boldsymbol{\Gamma} \exp \left(\boldsymbol{u}_i / \varepsilon \right) \right) \rangle$$
, $\boldsymbol{u}_i \in \mathbb{R}^N$, for all $i \in [n]$,

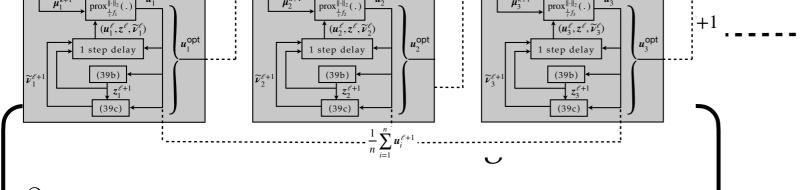
Then the following Euclidean ADMM solves (\bigvee)

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

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

$$\widetilde{oldsymbol{
u}}_i^{\ell+1} = \widetilde{oldsymbol{
u}}_i^\ell + ig(oldsymbol{u}_i^{\ell+1} - oldsymbol{z}_i^{\ell+1}ig)$$





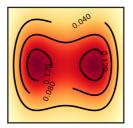
$$\frac{\partial \mu}{\partial t} = \nabla \cdot (\mu \nabla V) + \beta^{-1} \Delta \mu$$

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

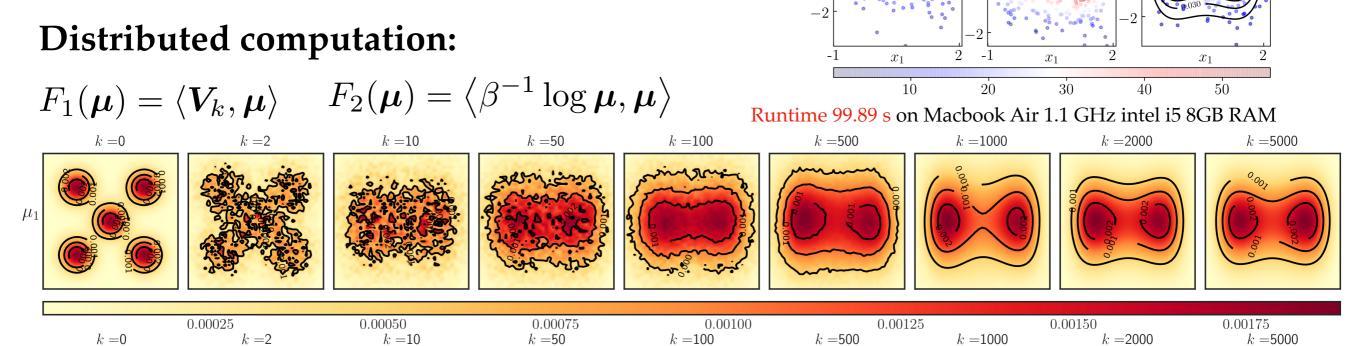
$$\mu_{\infty} \propto \exp(-\beta V(x_1, x_2)) dx_1 dx_2$$

0.00025

0.00050



0.00075



0.00100

0.00125

0.00150

0.00175

-2.5

Caluya and Halder, IEEE Trans. Automatic Control, 2019

 $\rho_{\text{\infty analytical}} = \frac{1}{Z} \exp\left(-\beta \psi(x_1, x_2)\right)$ • • ρ_{proximal}

Experiment #2

Aggregation-drift-diffusion nonlinear PDE

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

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

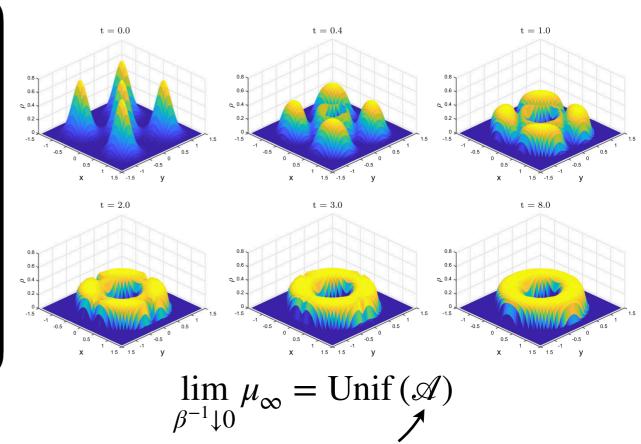
$$V(oldsymbol{x}) = -rac{1}{4} ext{ln} \, \|oldsymbol{x}\|_2$$

Distributed computation:

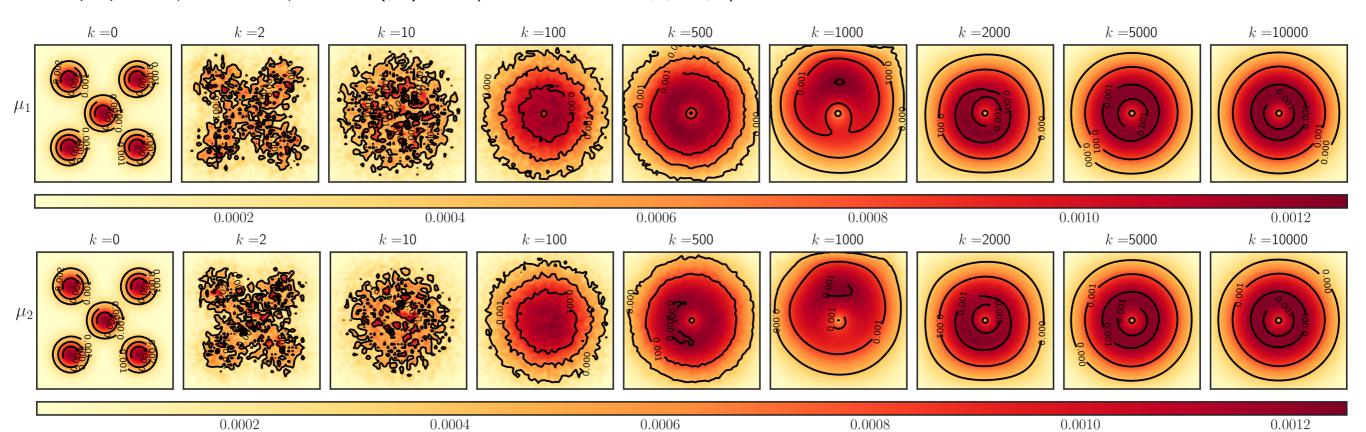
$$F_1(oldsymbol{\mu}) = \langle oldsymbol{U}_k oldsymbol{\mu}, oldsymbol{\mu}
angle \quad F_2(oldsymbol{\mu}) = \langle oldsymbol{V}_k + eta^{-1} \log oldsymbol{\mu}, oldsymbol{\mu}
angle$$

Centralized computation:

Carrillo, Craig, Wang and Wei, FOCM, 2021



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



Experiment #2 (contd.)

Aggregation-drift-diffusion nonlinear PDE

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

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

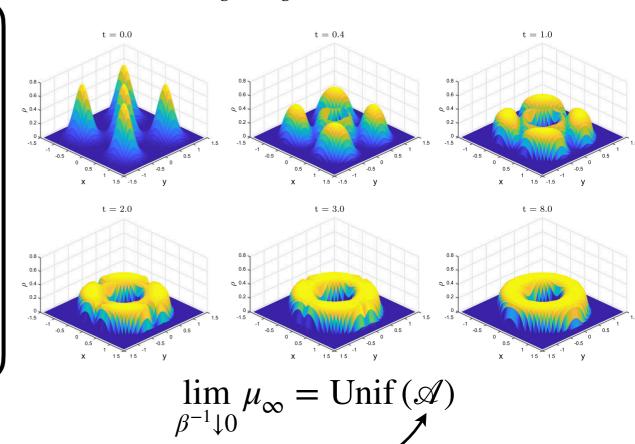
$$V(oldsymbol{x}) = -rac{1}{4} ext{ln} \, \|oldsymbol{x}\|_2$$

Distributed computation:

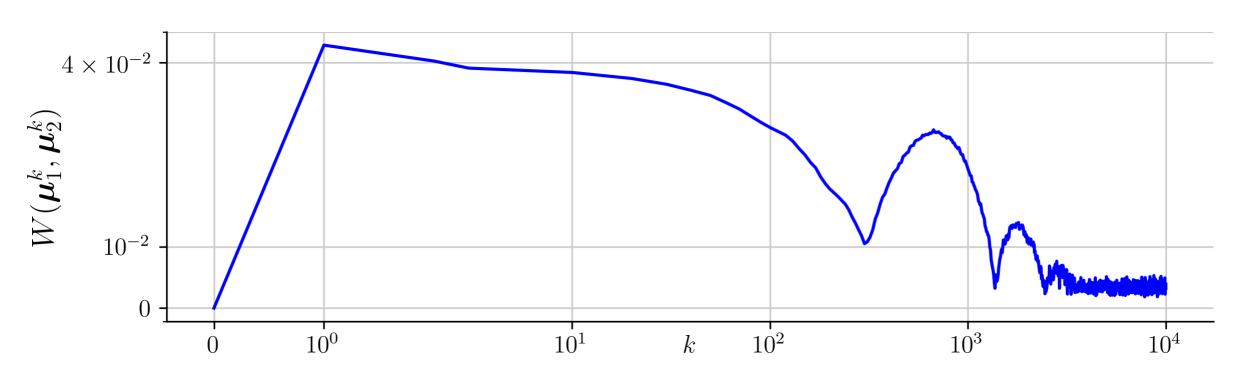
$$F_1(oldsymbol{\mu}) = \langle oldsymbol{U}_k oldsymbol{\mu}, oldsymbol{\mu}
angle \quad F_2(oldsymbol{\mu}) = \langle oldsymbol{V}_k + eta^{-1} \log oldsymbol{\mu}, oldsymbol{\mu}
angle
angle$$

Centralized computation:

Carrillo, Craig, Wang and Wei, FOCM, 2021



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



37

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	$F_1(oldsymbol{\mu}) = \left\langle oldsymbol{V}_k + eta^{-1}oldsymbol{\mu}, oldsymbol{\mu} ight angle, \ F_2(oldsymbol{\mu}) = \left\langle oldsymbol{U}_koldsymbol{\mu}^k, oldsymbol{\mu} ight angle,$	Splitting case #1 4×10^{-2} 4×10^{-3} 10^{-3} 10^{0} 10^{1} k 10^{2} 10^{3} 10^{4}
#2	$F_1(oldsymbol{\mu}) = \langle oldsymbol{U}_k oldsymbol{\mu}^k + eta^{-1} oldsymbol{\mu}, oldsymbol{\mu} angle, \ F_2(oldsymbol{\mu}) = \langle oldsymbol{V}_k, oldsymbol{\mu} angle$	Splitting case #2 4×10^{-2} 4×10^{-3} 10^{0} 10^{1} 10^{1} 10^{2} 10^{3} 10^{4}
#3	$F_1(oldsymbol{\mu}) = \langle oldsymbol{U}_k oldsymbol{\mu}^k + oldsymbol{V}_k, oldsymbol{\mu} angle, \ F_2(oldsymbol{\mu}) = \left\langle eta^{-1} oldsymbol{\mu}, oldsymbol{\mu} ight angle$	4×10^{-2} 4×10^{-3} 10^{0} 10^{1} k 10^{1} k 10^{2} 10^{3} 10^{4} 10^{1} k 10^{2} 10^{3} 10^{4}
#4	$F_1(oldsymbol{\mu}) = \langle oldsymbol{V}_k, oldsymbol{\mu} angle, \ F_2(oldsymbol{\mu}) = \langle oldsymbol{U}_k oldsymbol{\mu}^k angle, \ F_3(oldsymbol{\mu}) = \langle eta^{-1} oldsymbol{\mu}, oldsymbol{\mu} angle angle$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$

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	$F_1(m{\mu}) = \left< m{V}_k + eta^{-1}m{\mu}, m{\mu} \right>,$ $F_2(m{\mu}) = \left< m{U}_km{\mu}^k, m{\mu} \right>$ av. runtime = 294.06 s	Splitting case #1 4×10^{-2} 4×10^{-3} 10^{0} 10^{1} 10^{1} 10^{2} 10^{3} 10^{4}
#2	$F_1(m{\mu}) = \langle m{U}_k m{\mu}^k + eta^{-1} m{\mu}, m{\mu} angle, \ F_2(m{\mu}) = \langle m{V}_k, m{\mu} angle $ av. runtime = 285.32 s	Splitting case #2 $ \begin{array}{c} & & \\ & \\ & \\ & \\ & \\ & \\ & \\ & \\ & \\ &$
#3	$F_1(m{\mu}) = \langle m{U}_k m{\mu}^k + m{V}_k, m{\mu} angle, \ F_2(m{\mu}) = \langle eta^{-1} m{\mu}, m{\mu} angle$	$ \begin{array}{c ccccccccccccccccccccccccccccccccccc$
#4	av. runtime = 289.87 s $F_1(\boldsymbol{\mu}) = \langle \boldsymbol{V}_k, \boldsymbol{\mu} \rangle,$ $F_2(\boldsymbol{\mu}) = \langle \boldsymbol{U}_k \boldsymbol{\mu}^k \rangle,$ $F_3(\boldsymbol{\mu}) = \langle \beta^{-1} \boldsymbol{\mu}, \boldsymbol{\mu} \rangle$	$\begin{array}{c ccccccccccccccccccccccccccccccccccc$
	av. runtime = 108.99 s	$0 = 10^{0}$ 10^{1} $k = 10^{2}$ 10^{3} 10^{4}

Summary

Distributed computation for measure-valued optimization

Realizes measure-valued operator splitting

Inner ADMM minimizer #3 Central Processor Inner ADMM minimizer #1 Inner ADMM minimizer #2 Distributed Processor #1 Distributed Processor #2 Distributed Processor #3 Outer layer Outer layer Outer layer **ADMM ADMM ADMM** $\downarrow \mu_1^{k+1}$ Inner layer Inner layer Inner layer **ADMM ADMM ADMM**

Takes advantage of the existing proximal and JKO type algorithms

Ongoing

Convergence guarantees for the overall scheme

High dimensional case studies

Thank You