Sampling from Wasserstein barycenters

Workshop on Dynamics and Discretization: PDEs, Sampling, and Optimization

Thibaut Le Gouic
Joint work with **Chiheb Daaloul**, **Magali Tournus** and **Jacques Liandrat**
October 28th, 2021

École Centrale de Marseille, Institut de Mathématiques de Marseille

Averaging

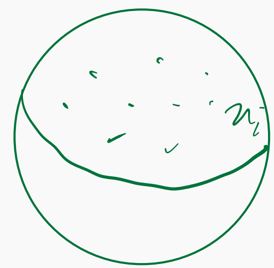
On $\mathbf{R}^d$:

$$\frac{1}{n}\sum x_i = \arg\min_x \frac{1}{n}\sum \|x_i - x\|^2$$

On the sphere:

$$\bar{x} = \frac{1}{n}\sum d(x_i, x)^2$$

(S, d)



Definition (Wasserstein distance)

For two measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu_0, \mu_1) = \inf_{\gamma \in \Gamma(\mu_0, \mu_1)} \int \|x - y\|^2 d\gamma(x, y).$$

- $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a geodesic space
- when $\mu_0 \ll \lambda, \quad \gamma^\star = (\mathrm{id}, T^{\mu_0 \to \mu_1})_\# \mu_0$
- it is positively curved

Wasserstein barycenter

Definition (Barycenter a.k.a. Fréchet mean - Agueh and Carlier 2011)

$\mu_1, \cdots, \mu_n$ probability measures on $\mathbf{R}^d$ with associated weights $\lambda_1, \cdots, \lambda_n$. Their barycenter is

$$\mu^\star \in \operatorname{argmin}_{\mu \in \mathcal{P}_2(\mathbf{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i). = F(\mu)$$

- Barycenter always exists
- Not always unique

Wasserstein barycenter

Definition (Barycenter a.k.a. Fréchet mean - Agueh and Carlier 2011)

$\mu_1, \cdots, \mu_n$ probability measures on $\mathbb{R}^d$ with associated weights $\lambda_1, \cdots, \lambda_n$. Their barycenter is

$$\mu^\star \in \mathrm{argmin}_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i).$$

- Barycenter always exists
- Not always unique

How to compute $\mu^\star$?

Finite support vs sampling

Most studied numerical setting to compute Wasserstein barycenters

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_{i,j}}, \quad i = 1, \cdots, n$$

This is NP-hard in (N, n, d) [Altschuler and Boix-Adsera 2021].

Finite support vs sampling

Most studied numerical setting to compute Wasserstein barycenters

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_{i,j}}, \quad i = 1, \cdots, n$$

This is NP-hard in (N, n, d) [Altschuler and Boix-Adsera 2021].

$\nabla \log \mu_i$

When the $x_{i,j}$ are drawn from a sampling procedure, can we do better?

i.e. how to sample *directly* from the barycenter $\mu^\star$ of $(\mu_i)_{i=1,\cdots,n}$ with weights $(\lambda_i)_{i=1,\cdots,n}$?

Finite support vs sampling

Most studied numerical setting to compute Wasserstein barycenters

$$\mu_i = \frac{1}{N} \sum_{j=1}^{N} \delta_{x_{i,j}}, \quad i = 1, \cdots, n$$

This is NP-hard in (N, n, d) [Altschuler and Boix-Adsera 2021].

When the $x_{i,j}$ are drawn from a sampling procedure, can we do better?

i.e. how to sample *directly* from the barycenter $\mu^\star$ of $(\mu_i)_{i=1,\cdots,n}$ with weights $(\lambda_i)_{i=1,\cdots,n}$?

$\rightarrow$ mutimarginal problem

Theorem (Agueh and Carlier 2011)

$$\inf_{\mu} \sum_{i=1}^{n} \lambda_i W_2^2(\mu_i, \mu) = \inf_{\gamma \in \Gamma(\mu_1, \cdots, \mu_n)} \int c \, d\gamma$$

$F(\mu)$

with

$$c(x_1, \cdots, x_n) = \sum_{i=1}^{n} \lambda_i \|x_i - \sum_{j=1}^{n} \lambda_j x_j\|^2.$$

And moreover

$$\mu^\star = ((x_1, \cdots, x_n) \mapsto \sum_{i=1}^{n} \lambda_i x_i)_{\#} \gamma^\star.$$

Relaxed problem

To sample using a flow gradient of the multimarginal formulation, we penalize to account for the constraints. For $\alpha > 0$, let

$$F_\alpha := \gamma \mapsto \int c \mathrm{d}\gamma + \alpha \sum_{i=1}^{n} \lambda_i \chi^2(\gamma_i | \mu_i),$$

where γ_i is the i-th marginal of γ.

Sampling as a Wasserstein gradient flow

$$F_\alpha := \gamma \mapsto \int c \, \mathrm{d}\gamma + \alpha \sum_{i=1}^{n} \lambda_i \, \chi^2(\gamma_i | \mu_i),$$

How to sample from the minimum of F_α?

Wasserstein gradient on $\mathcal{P}_2(\mathbf{R}^{d \times n})$.

$$\nabla_W F_\alpha(\gamma) := \nabla_x c + \alpha \sum_{i=1}^{n} \lambda_i \nabla_{x_i}(\gamma_i / \mu_i).$$

$$\mathbb{R}^{d \times n} \to \mathbb{R}^{d \times n}$$

$$F_\alpha := \gamma \mapsto \int c \mathrm{d}\gamma + \alpha \sum_{i=1}^n \lambda_i \chi^2(\gamma_i | \mu_i),$$

How to sample from the minimum of F_α?

Wasserstein gradient on $\mathcal{P}_2(\mathbf{R}^{d\times n})$:

$$\nabla_W F_\alpha(\gamma) := \nabla_x c + \alpha \sum_{i=1}^n \lambda_i \nabla_{x_i}(\gamma_i/\mu_i).$$

Since γ_i is unknown, replace $\nabla(\gamma_i/\mu_i)$ with kernelized version

$$\int \nabla(\gamma_i/\mu_i)(y) K(x,y) \mathrm{d}\mu_i(y)$$

$$=$$

$$-\underbrace{\int \nabla_y K(x,y) \mathrm{d}\gamma_i(y)}_{\approx \frac{1}{n}\sum_{j=1}^n \nabla_y K(x, X_{i,j})} - \underbrace{\int \nabla \log(\mu_i)(y) K(x,y) \mathrm{d}\gamma_i(y)}_{\approx \frac{1}{n}\sum_{j=1}^n \nabla \log(\mu_i)(X_{i,j}) K(x, X_{i,j})}.$$

This is the Stein Variational Gradient Descent (SVGD).Liu and Wang 2016

Chewi, **TLG**, Lu, Maunu, and Rigollet 2020

Wasserstein flow

$$\dot{X}^t = -\nabla_x c(X^t) + \alpha \sum_{i=1}^{n} \lambda_i \nabla_{x_i}(\gamma_i/\mu_i)(X_i^t).$$

Implementation: choose kernel K and step size $h > 0$, draw N particles $X_1^0, \cdots, X_N^0$ in $(\mathbf{R}^d)^n$ and iterate

$$X_{i,j}^{t+1} - X_{i,j}^t = -h \underbrace{\nabla c(X_{1,j}^t, \cdots, X_{n,j}^t)}_{\text{interaction between marginals of particle } j}$$

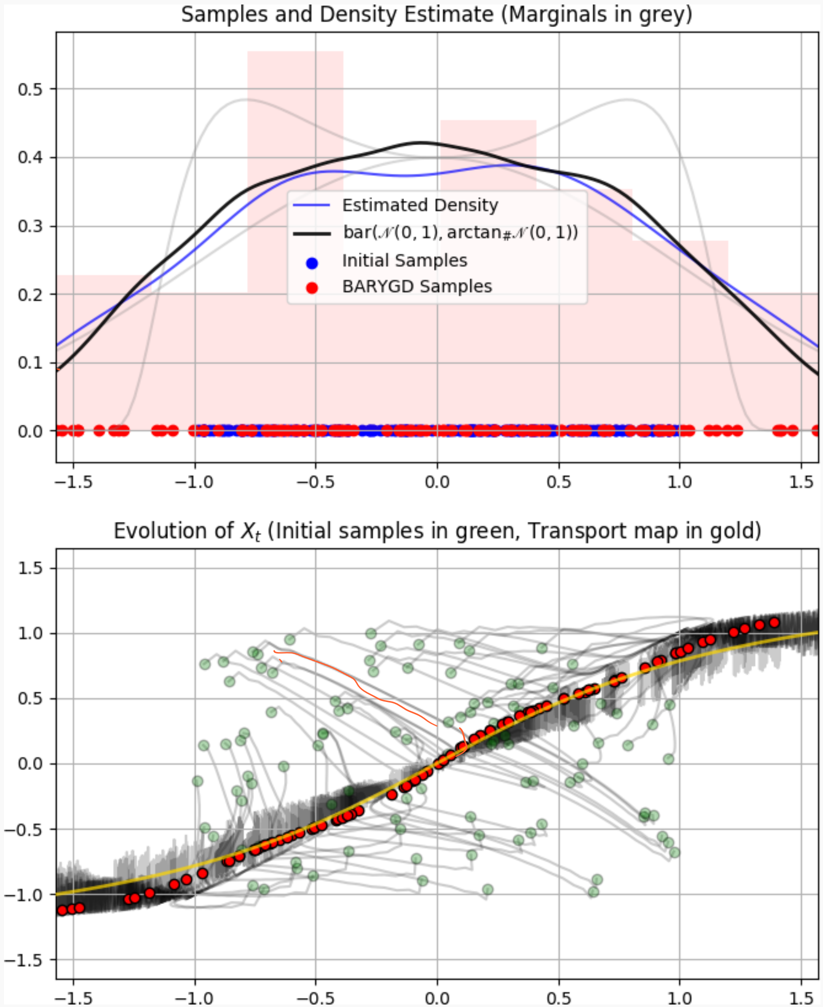
$$+h\alpha \sum_{i=1}^{n} \lambda_i \left(\underbrace{\frac{1}{N}\sum_{k=1}^{N} \nabla_y K(X_{i,j}^t, X_{i,k}^t) + \frac{1}{N}\sum_{k=1}^{N} \nabla \log(\mu_i)(X_{i,k}) K(X_{i,j}^t, X_{i,k}^t)}_{\text{interaction between the same marginal } i \text{ of all particles}} \right).$$

Numerical experiments

d = 1

n = 2

N = 50

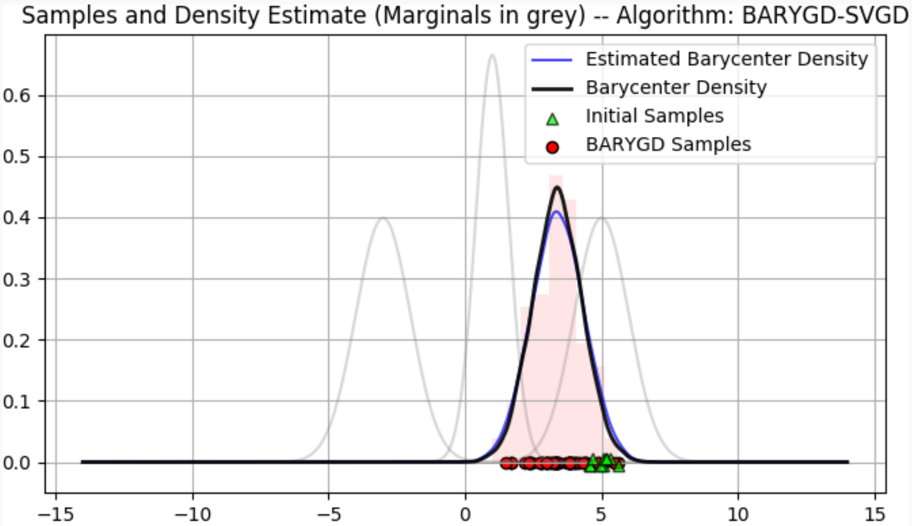


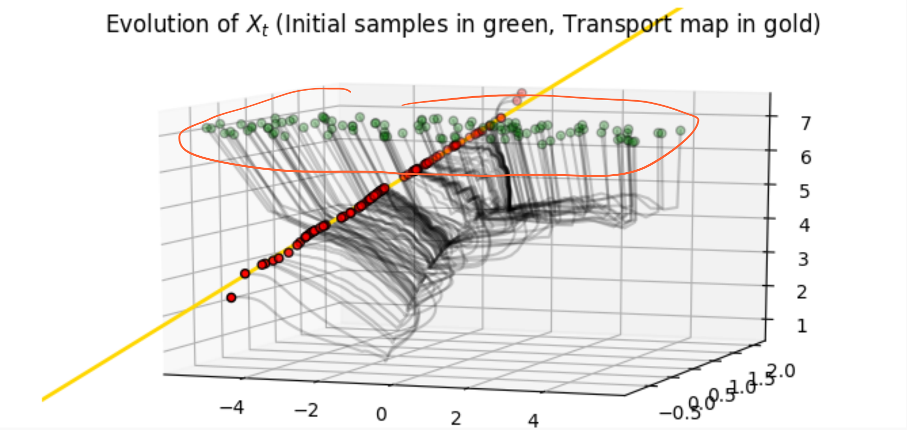
Numerical experiments

$d = 1$

$n = 3$

$N = 50$





Variance inequality

$$F_\alpha := \gamma \mapsto \int c \mathrm{d}\gamma + \alpha \sum_{i=1}^n \lambda_i \, \chi^2(\gamma_i | \mu_i),$$

Denoting $\gamma_\alpha^\star$ the minimizer of F_α, is the associated barycenter

$$\mu_\alpha^\star := (x \mapsto \sum_{i=1}^n \lambda_i x_i)_{\#} \gamma_\alpha^\star$$

close the the true barycenter $\mu^\star$?

Variance inequality

$$F_\alpha := \gamma \mapsto \int c \mathrm{d}\gamma + \alpha \sum_{i=1}^{n} \lambda_i \, \chi^2(\gamma_i | \mu_i),$$

Denoting $\gamma_\alpha^\star$ the minimizer of F_α, is the associated barycenter

$$\mu_\alpha^\star := (x \mapsto \sum_{i=1}^{n} \lambda_i x_i)_\# \gamma_\alpha^\star$$

close the the true barycenter $\mu^\star$?

Theorem (Uniqueness Agueh and Carlier 2011; Daaloul, TLG, Liandrat, and Tournus 2021**)**

If one the μ_i's is absolutely continuous w.r.t. then $\mu^\star$ and $\mu_\alpha^\star$ are unique.

Lebesgue

$$F_\alpha := \gamma \mapsto \int c \mathrm{d}\gamma + \alpha \sum_{i=1}^{n} \lambda_i \chi^2(\gamma_i | \mu_i),$$

Denoting $\gamma_\alpha^\star$ the minimizer of F_α, is the associated barycenter

$$\mu_\alpha^\star := (x \mapsto \sum_{i=1}^{n} \lambda_i x_i)_{\#}\gamma_\alpha^\star$$

close the the true barycenter $\mu^\star$?

Assumption (Variance inequality)

There exists $k > 0$ such that

$$\underbrace{\sum_{i=1}^{n} \lambda_i W_2^2(\mu, \mu_i)}_{F(\mu)} - \underbrace{\sum_{i=1}^{n} \lambda_i W_2^2(\mu^\star, \mu_i)}_{F(\mu^\star)} \geq k W_2^2(\mu^\star, \mu)$$

- This is also known as *quadratic growth* in the optimization literature.
- Implies uniqueness of the barycenter.

Note that this is always true for $k = 0$.

Variance inequality controls relaxation

Theorem (Daaloul, TLG, Liandrat, and Tournus 2021)

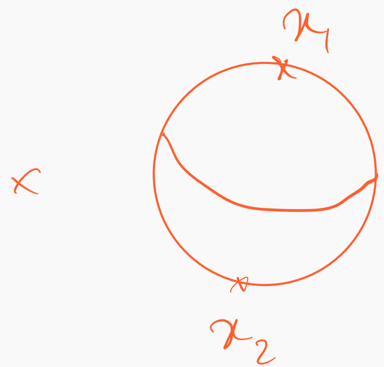
Suppose each $\mu_1, \ldots, \mu_n$ satisfy a Poincaré inequality with constant C_P and that $\sum \lambda_i \delta_{\mu_i}$ satisfies a variance inequality with constant k. Denote

$$\mu_\alpha^\star = ((x_1, \cdots, x_n) \mapsto \lambda_i x_i)_\# \gamma_\alpha^\star.$$

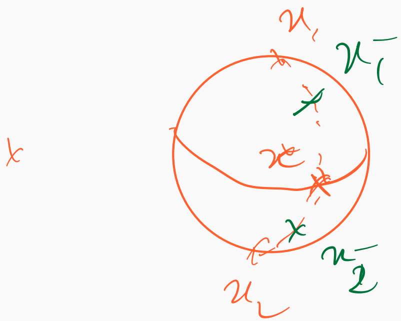
Then,

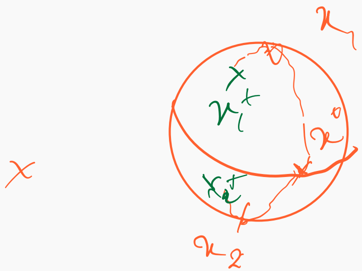
$$k\, W_2^2(\mu^\star, \mu_\alpha^\star) \leq \frac{16\, C_P}{\alpha} \int c\, d\gamma^\star.$$

When does it hold?



non unique





x^* is not barycenter of x_1^+ and x_2^+

Theorem (Variance inequality - Ahidar-Coutrix, TLG, and Paris 2020)

Let $x^\star$ be the barycenter of $x_1, \cdots, x_n$ with weights $\lambda_1, \cdots, \lambda_n$ on a positively curved geodesic space. Denote

$$x_i^+ = x^\star + (1+\lambda)(x_i - x^\star).$$

If $x^\star$ is still the barycenter of $x_1^+, \cdots, x_n^+$, then $x_1, \cdots, x_n$ satisfies a $\frac{\lambda}{1+\lambda}$-variance inequality.

Variance inequality

Theorem (Variance inequality - Ahidar-Coutrix, TLG, and Paris 2020)

Let $x^\star$ be the barycenter of $x_1, \cdots, x_n$ with weights $\lambda_1, \cdots, \lambda_n$ on a positively curved geodesic space. Denote

$$x_i^+ = x^\star + (1+\lambda)(x_i - x^\star).$$

If $x^\star$ is still the barycenter of $x_1^+, \cdots, x_n^+$, then $x_1, \cdots, x_n$ satisfies a $\frac{\lambda}{1+\lambda}$-variance inequality.

What does it mean in the Wasserstein space?

Theorem (Variance inequality - Ahidar-Coutrix, TLG, and Paris 2020)

Let $x^\star$ be the barycenter of $x_1, \cdots, x_n$ with weights $\lambda_1, \cdots, \lambda_n$ on a positively curved geodesic space. Denote

$$x_i^+ = x^\star + (1+\lambda)(x_i - x^\star).$$

If $x^\star$ is still the barycenter of $x_1^+, \cdots, x_n^+$, then $x_1, \cdots, x_n$ satisfies a $\frac{\lambda}{1+\lambda}$-variance inequality.

What does it mean in the Wasserstein space?

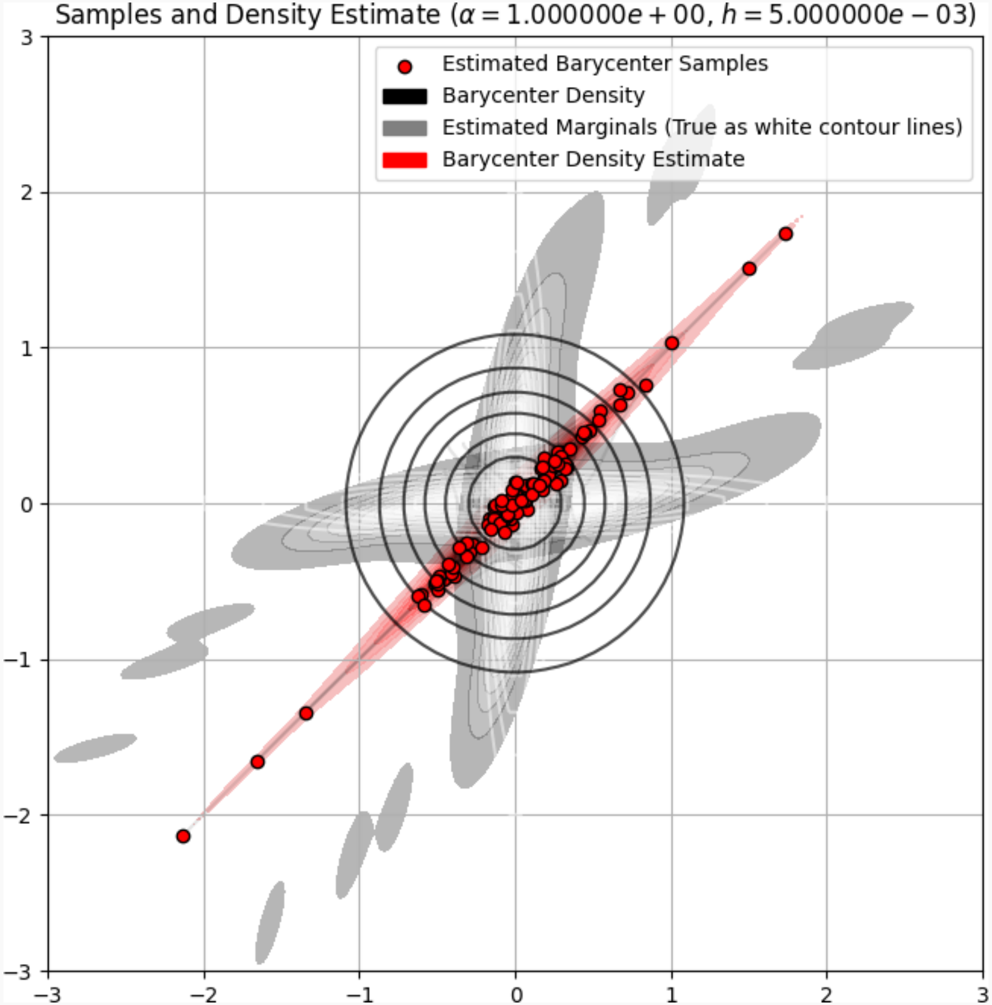
Theorem (Variance inequality in $\mathcal{P}_2(\mathbb{R}^d)$ — Chewi, Maunu, Rigollet, and Stromme 2020)

Suppose the support of $\mu^\star \ll \lambda$ is $\mathbb{R}^d$. If for all i, the Kantorovich potential $\phi^{\mu^\star \to \mu_i}$ is α_i-strongly convex, then a k-variance inequality holds with

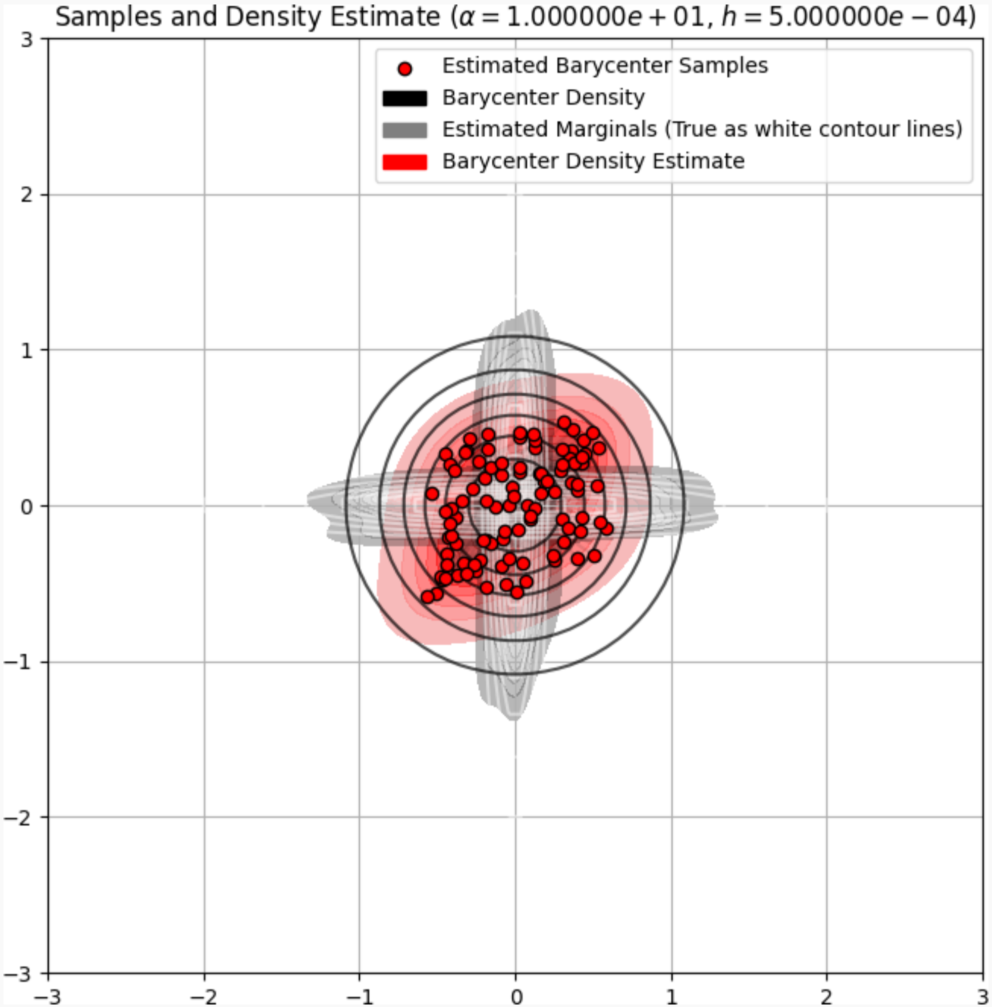
$$k = \sum_{i=1}^{n} \lambda_i \alpha_i.$$

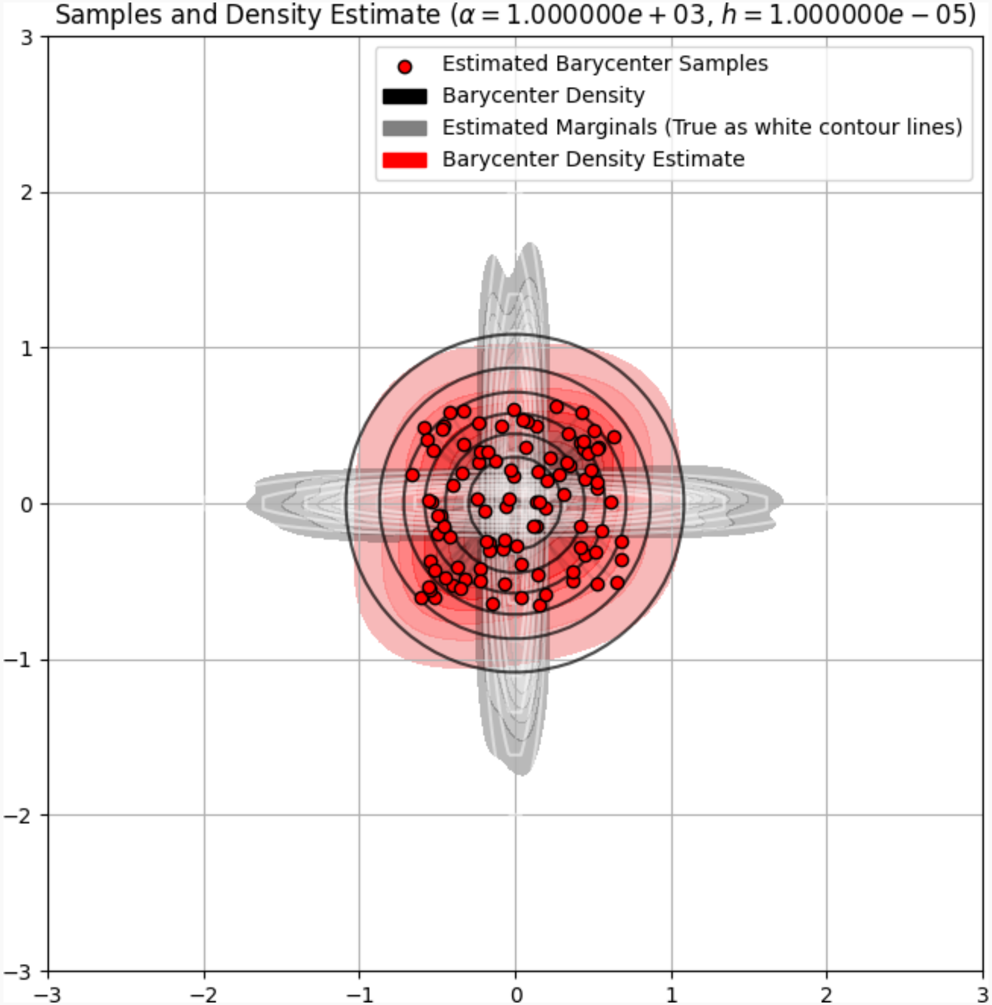
Numerical experiments

$\Sigma = \begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix}$



Numerical experiments

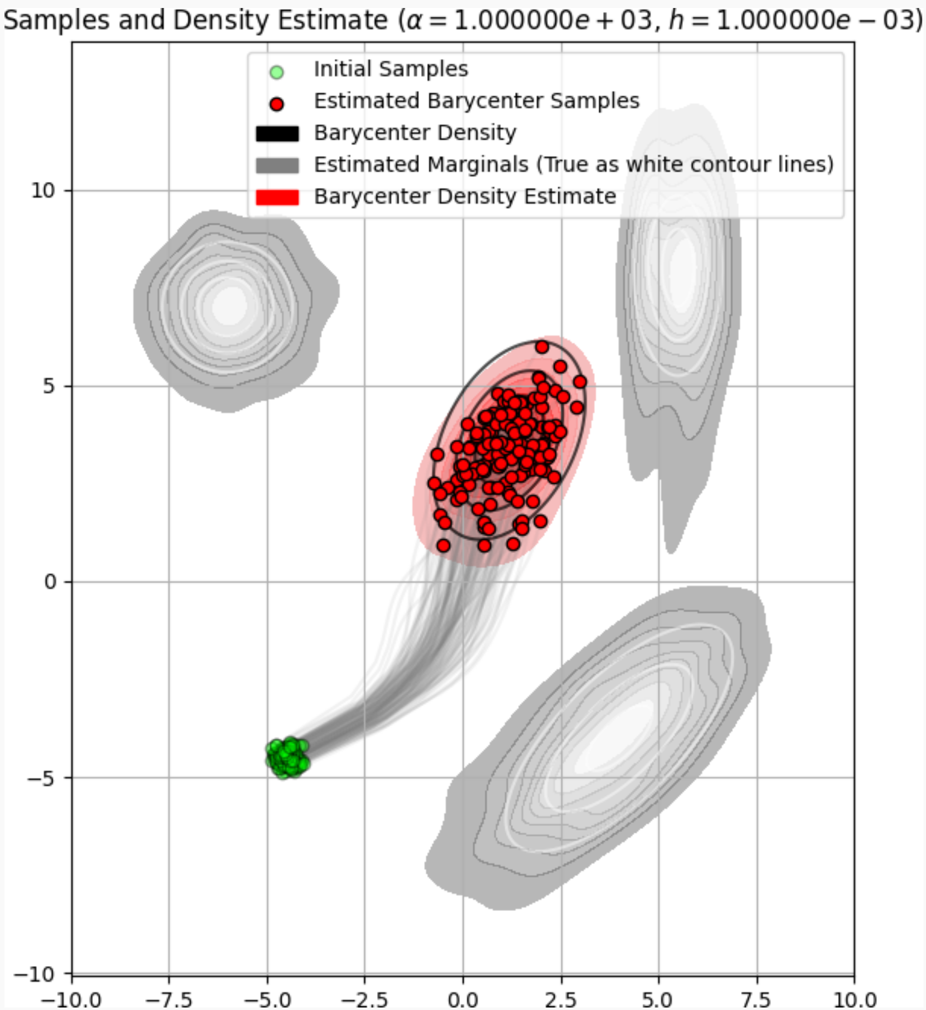



Samples and Density Estimate (α = 1.000000e + 03, h = 1.000000e − 05)
Estimated Barycenter Samples
Barycenter Density
Estimated Marginals (True as white contour lines)
Barycenter Density Estimate
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Closing remarks

Open questions:

- What rate of convergence?
- What dependence on the dimension?
- Other costs for relaxed/unbalanced multimarginal problem?


Samples and Density Estimate ($\alpha = 1.000000e + 03$, $h = 1.000000e - 03$)
Initial Samples
Estimated Barycenter Samples
Barycenter Density
Estimated Marginals (True as white contour lines)
Barycenter Density Estimate
10
5
0
−5
−10
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

References

Agueh, Martial and Guillaume Carlier (2011). “Barycenters in the Wasserstein space”. In: *SIAM Journal on Mathematical Analysis* 43.2, pp. 904–924.

Ahidar-Coutrix, Adil, **TLG**, and Quentin Paris (2020). “Convergence rates for empirical barycenters in metric spaces: curvature, convexity and extendable geodesics”. In: *Probability theory and related fields* 177.1, pp. 323–368.

Altschuler, Jason M and Enric Boix-Adsera (2021). “Wasserstein barycenters are NP-hard to compute”. In: *arXiv preprint arXiv:2101.01100*.

Chewi, Sinho, **TLG**, Chen Lu, Tyler Maunu, and Philippe Rigollet (2020). “SVGD as a kernelized Wasserstein gradient flow of the chi-squared divergence”. In: *arXiv preprint arXiv:2006.02509*.

Chewi, Sinho, Tyler Maunu, Philippe Rigollet, and Austin J Stromme (2020). “Gradient descent algorithms for Bures-Wasserstein barycenters”. In: *Conference on Learning Theory*. PMLR, pp. 1276–1304.

Daaloul, Chiheb, **TLG**, Jacques Liandrat, and Magali Tournus (2021). "Sampling From the Wasserstein Barycenter". In: *arXiv preprint arXiv:2105.01706*.

Liu, Qiang and Dilin Wang (2016). "Stein variational gradient descent: A general purpose bayesian inference algorithm". In: *arXiv preprint arXiv:1608.04471*.