# Statistical aspects of stochastic algorithms for entropic optimal transportation between probability measures

**Jérémie Bigot**

Institut de Mathématiques de Bordeaux
Equipe Image, Optimisation et Probabilités (IOP)

**Université de Bordeaux**

Joint work with **Bernard Bercu** (IMB, Bordeaux)

**Statistical modeling for shapes and imaging**

The Mathematics of Imaging, IHP, March 2019

# An example of a ressource allocation problem

**Data at hand [1] :**

- locations of Police stations in Chicago
- spatial locations of reported incidents of crime (with the exception of murders) in Chicago in 2014

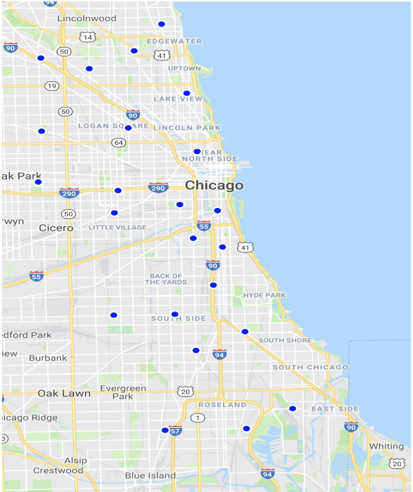
**Questions (of interest ?) :**

- given the location of a crime, which Police station should intervene ?
- how updating the answer in an "online fashion" along the year ?

1. Open Data from Chicago : `https://data.cityofchicago.org`

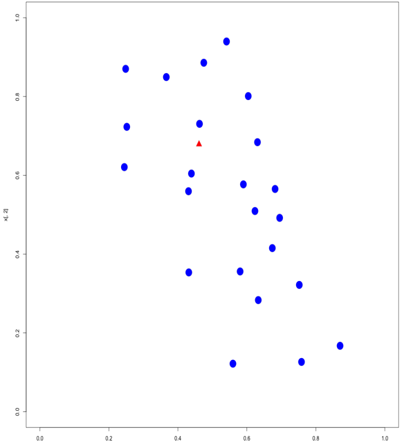
# An example of a ressource allocation problem

Locations $y_1, \ldots, y_J$ of Police stations in Chicago



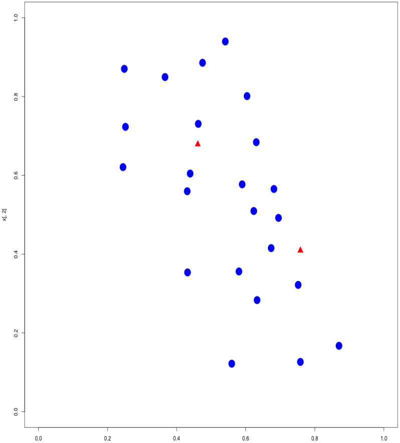
# An example of a ressource allocation problem

Spatial location $X_1$ of the **first** reported incident of crime in Chicago in the year 2014



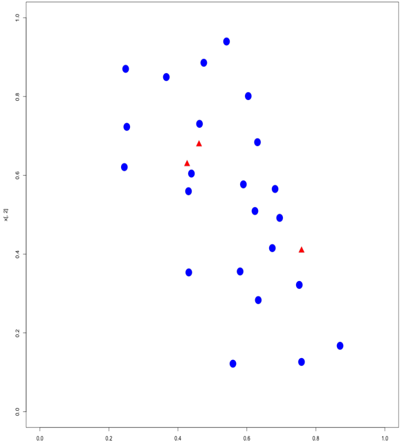
# An example of a ressource allocation problem

Spatial locations $X_1, X_2$ of reported incidents of crime in Chicago in **chronological order**



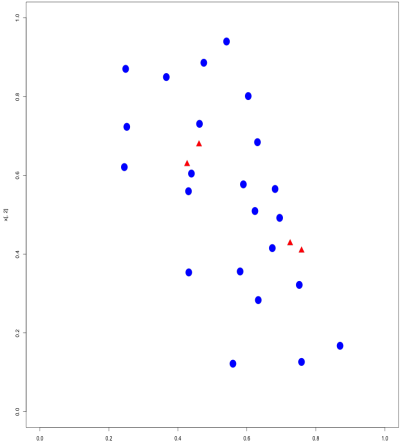
# An example of a ressource allocation problem

Spatial locations $X_1, X_2, X_3$ of reported incidents of crime in Chicago in **chronological order**



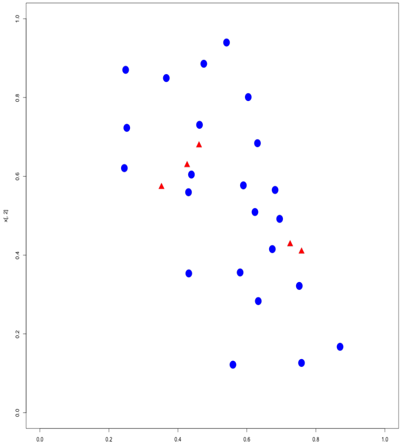
# An example of a ressource allocation problem

Spatial locations $X_1, \ldots, X_4$ of reported incidents of crime in Chicago in **chronological order**



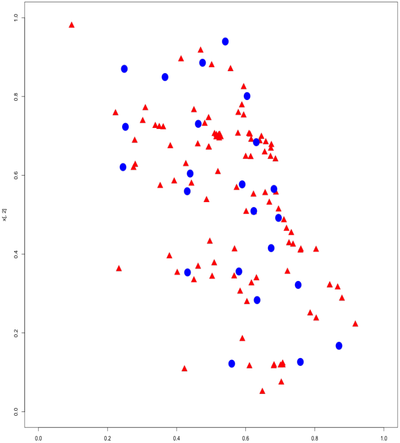
# An example of a ressource allocation problem

Spatial locations $X_1, \ldots, X_5$ of reported incidents of crime in Chicago in **chronological order**



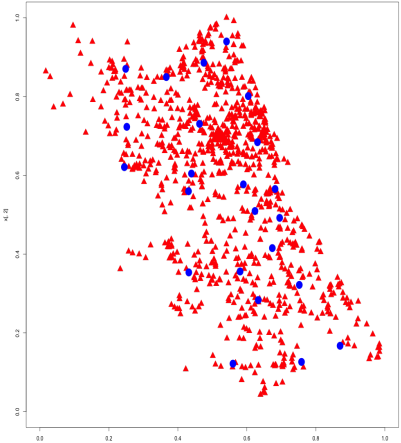
# An example of a ressource allocation problem

Spatial locations of reported incidents of crime in Chicago in **chronological order** (first 100)



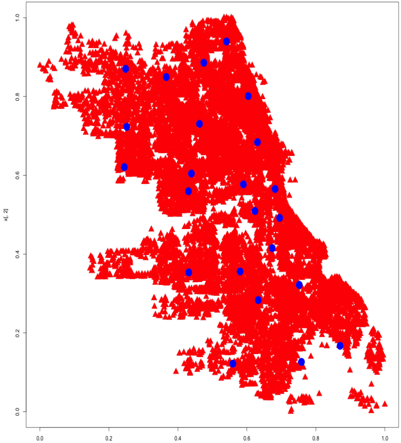
# An example of a ressource allocation problem

Spatial locations of reported incidents of crime in Chicago in **chronological order** (first 1000)



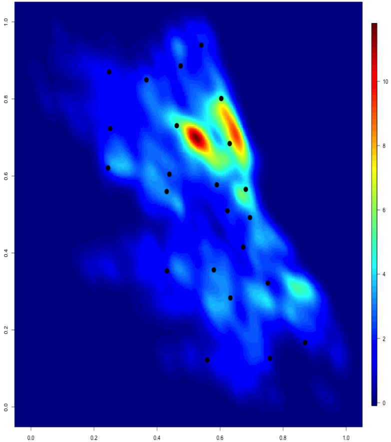
# An example of a ressource allocation problem

Spatial locations $X_1, \ldots, X_N$ of reported incidents of crime in Chicago in **chronological order** (total $N = 16104$)



# An example of a ressource allocation problem

Heat map (kernel density estimation) of spatial locations of reported incidents of crime in Chicago in 2014



# Statistical approach to ressource allocation

**Modeling assumptions :**

- spatial locations of reported incidents of crime : a sequence of iid random variables
$$X_1, \ldots, X_n$$
sampled from an **unknown** probability measure $\mu$ with support $\mathcal{X} \subset \mathbb{R}^2$

- locations of Police station : a **known and discrete** probability measure
$$\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$$
where
  - $y_j \in \mathbb{R}^2$ represent the spatial location of the $j$-th Police station
  - $\nu_j$ is a positive weight representing the “capacity” of each Police station (we took $\nu_j = 1/J$ that is uniform weights)

# Statistical approach to ressource allocation

**Point of view in this talk :** ressource allocation can be solved by finding an optimal transportation map

$$T : \mathcal{X} \to \{y_1, \dots, y_J\}$$

which pushes forward $\mu$ onto $\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$ (notation : $T\#\mu = \nu$), with respect to a given cost function, e.g. a distance on $\mathcal{X}$

$$c(x, y) = \|x - y\|_{\ell_p} = \left( \sum_{k=1}^{d} (x_k - y_k)^p \right)^{1/p}, \quad x, y \in \mathbb{R}^d \text{ (here } d = 2)$$

**Question :** how doing on-line estimation of such a map using the observations $X_1, \dots, X_n \sim_{iid} \mu$ ?

# Optimal transport between probability measures

- Let $T : \mathcal{X} \to \{y_1, \ldots, y_J\}$ such that $T\#\mu = \nu$
- Let $\Pi(\mu, \nu)$ be the set of probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals $\mu$ and $\nu$

**Definition**

*The optimal transport problem between $\mu$ and $\nu$ is*

$$W_0(\mu, \nu) = \min_{T \,:\, T\#\mu = \nu} \int_{\mathcal{X}} c(x, T(x)) d\mu(x), \text{ (Monge's formulation)}$$

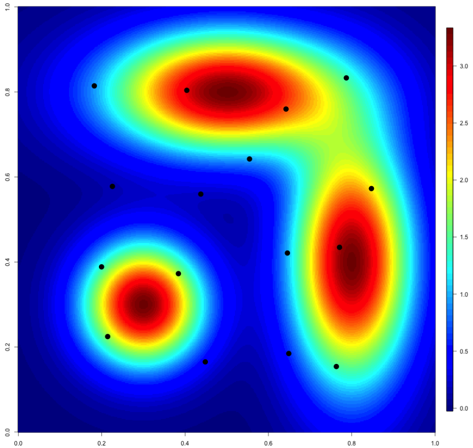
*or*

$$W_0(\mu, \nu) = \min_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\pi(x, y), \text{ (Kantorovich's formulation)}$$

*where $c(x, y)$ is the cost function of moving mass from $x$ to $y$.*

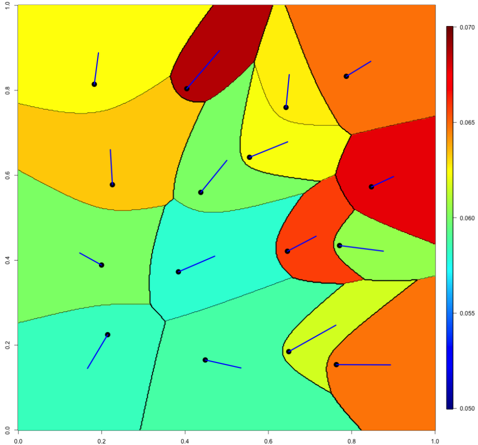
# An example of semi-discrete optimal transport

Optimal transport of an absolutely continuous measure $\mu$ onto a discrete measure $\nu$ (black dots)



# An example of semi-discrete optimal transport

Optimal transport of $\mu$ onto the discrete measure $\nu$ (black dots) - Optimal map $T$ for the Euclidean cost $c(x,y) = \|x-y\|_{\ell_2}$



# Semi-discrete optimal transport

Unicity of an optimal mapping $T : \mathrm{supp}(\mu) \to \{y_1, \ldots, y_J\}$ such that $T\#\mu = \nu$ given, for all $1 \leq j \leq J$, by [1]

$$T^{-1}(y_j) = \left\{ x \in \mathrm{supp}(\mu) \;:\; c(x, y_j) - v^*_{j,0} \leq c(x, y_k) - v^*_{k,0} \text{ for all } 1 \leq k \leq J \right\}$$

where $v^*_0 \in \mathbb{R}^J$ is any maximizer of the un-regularized semi-dual problem of the Kantorovich's formulation of OT.

The sets $\{T^{-1}(y_j)\}$ are the so-called Laguerre cells (important concept from computational geometry).

1. Mérigot (2018), Cuturi and Peyré (2017)

# Optimal transport between probability measures

**Problem :** computational cost of optimal transport for data analysis [1]

**Case of discrete measures :** if

$$\mu = \sum_{i=1}^{K} \mu_i \delta_{x_i} \text{ and } \nu = \sum_{j=1}^{K} \nu_j \delta_{y_j}$$

then the cost to evaluate $W_0(\mu, \nu)$ (linear program) is generally

$$\mathcal{O}(K^3 \log K)$$

1. See the recent book by Cuturi & Peyré (2018)

# Regularized optimal transport

**Definition (Cuturi (2013))**

*Let $\mu$ and $\nu$ be any probability measures supported on $\mathcal{X}$. Then, the regularized optimal transport problem between $\mu$ and $\nu$ is*

$$W_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} c(x,y)d\pi(x,y) + \varepsilon KL(\pi|\mu\otimes\nu),$$

*where $\epsilon > 0$ (regularization parameter) and*

$$KL(\pi|\xi) = \int_{\mathcal{X}\times\mathcal{X}} \left(\log\left(\frac{d\pi}{d\xi}(x,y)\right) - 1\right)d\pi(x,y), \text{ with } \xi = \mu\otimes\nu.$$

**Case of discrete measures :** for $\epsilon > 0$

- Sinkhorn algorithm (iterative scheme) to compute $W_\varepsilon(\mu,\nu)$
- computational cost of $\mathcal{O}(K^2)$ at each iteration

# Stochastic optimal transport

## Proposition (Genevay, Cuturi, Peyré and Bach (2016))

*Let* $\mu$ *be any probability measure and* $\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j}$. *For* $\varepsilon > 0$, *solve the* **smooth concave maximization** *problem*

$$W_\varepsilon(\mu,\nu) = \max_{v \in \mathbb{R}^J} H_\varepsilon(v), \text{ where } \underbrace{H_\varepsilon(v) := \mathbb{E}[h_\varepsilon(X,v)]}_{\text{Stochastic optimization}}$$

*where* $X$ *is a random variable with distribution* $\mu$, *and for* $x \in \mathcal{X}$ *and* $v \in \mathbb{R}^J$,

$$h_\varepsilon(x,v) \quad = \quad \sum_{j=1}^{J} v_j \nu_j - \varepsilon \log\left( \sum_{j=1}^{J} \exp\left( \frac{v_j - c(x,y_j)}{\varepsilon} \right) \nu_j \right) - \varepsilon.$$

# Stochastic algorithm [1]

For fixed $\epsilon > 0$, Robbins-Monro algorithm to compute a minimizer

$$v^* := v_\varepsilon^* \in \arg\min_{v \in \mathbb{R}^J} \mathbb{E}[h_\varepsilon(X, v)]$$

Let $X_1, \ldots, X_n \sim_{iid} \mu$, choose $V_0 \in \mathbb{R}^J$ and a sequence $\gamma_{n+1}$ of steps with $\sum_{n=1}^{\infty} \gamma_n = +\infty$ and $\sum_{n=1}^{\infty} \gamma_n^2 < +\infty$ and do

$$\widehat{V}_{n+1} = \widehat{V}_n + \gamma_{n+1} \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n)$$

**Easy computation of gradients for $\epsilon > 0$** (smooth optimization)

$$\nabla_v h_\varepsilon(x, v) = \nu - \pi(x, v)$$

where $\pi(x, v) \in \mathbb{R}^J$ with

$$\pi_j(x, v) = \left( \sum_{k=1}^{J} \nu_k \exp\left( \frac{v_k - c(x, y_k)}{\varepsilon} \right) \right)^{-1} \nu_j \exp\left( \frac{v_j - c(x, y_j)}{\varepsilon} \right)$$

1. Genevay, Cuturi, Peyré and Bach (2016), Galerne, Leclaire, Rabin (2018)

# Contribution in our work [1]

**Main results** on the sequence $\widehat{V}_n$ : assume that the step $\gamma_n = \gamma/n$ where $\gamma > 0$ satisfies

$$\gamma > \frac{1}{2\rho^*}$$

where $\rho^*$ denotes the (second) smallest value of the Hessian matrix

$$-\nabla^2 H_\varepsilon(v) \text{ at } v = v^*,$$

or that $\gamma_n = \gamma/n^c$ where $\gamma > 0$ and $1/2 < c < 1$.

## Proposition

**Then**, $\lim_{n\to\infty} \widehat{V}_n = v^*$ *almost surely, and one has the asymptotic normality of*

$$\sqrt{n^c}\left(\widehat{V}_n - v^*\right)$$

*as* $n \to +\infty$.

1. Bercu, B. & Bigot, J. (2018) ArXiv :1812.09150

# Contribution in our work [1]

**Main results** on the sequence $\widehat{V}_n$ : assume that the step $\gamma_n = \gamma/n$ where $\gamma > 0$ satisfies

$$\gamma > \frac{1}{2\rho^*}$$

where $\rho^*$ denotes the (second) smallest value of the Hessian matrix

$$-\nabla^2 H_\varepsilon(v) \text{ at } v = v^*,$$

or that $\gamma_n = \gamma/n^c$ where $\gamma > 0$ and $1/2 < c < 1$.

**Interestingly**, one has that

$$-\nabla^2 H_\varepsilon(v^*) = \frac{1}{\varepsilon}\Big(\mathbb{E}\big[\pi(X, v^*)\pi(X, v^*)^T\big] - \mathrm{diag}(\nu)\Big)$$

which is not far from the covariance matrix of a multinomial distribution, implying that

$$\frac{1}{\varepsilon}\min_{1\leq j\leq J} \nu_j \leq \rho^* \leq \frac{1}{\varepsilon}, \text{ hence we took } \gamma = \frac{\varepsilon}{2\min_{1\leq j\leq J}\nu_j}$$

1. Bercu, B. & Bigot, J. (2018) ArXiv :1812.09150

# Contribution in our work [1]

**Main goal :** estimation of the Wasserstein functional $W_\varepsilon(\mu,\nu)$ based on $X_1,\ldots,X_n \sim_{iid} \mu$ and assuming that $\nu$ is known

**A simple recursive estimator :**

$$\widehat{W}_n = \frac{1}{n}\sum_{k=1}^{n} h_\varepsilon(X_k, \widehat{V}_{k-1}).$$

**Main results :** a.s. convergence of $\widehat{W}_n$ + asymptotic normality with same conditions for $\gamma_n$

$$\sqrt{n}\left(\widehat{W}_n - W_\varepsilon(\mu,\nu)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \sigma^2_\varepsilon(\mu,\nu)\right)$$

where the asymptotic variance $\sigma^2_\varepsilon(\mu,\nu)$ can also be estimated in a recursive manner

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n} h^2_\varepsilon(X_k, \widehat{V}_{k-1}) - \widehat{W}_n^2.$$

1. Bercu, B. & Bigot, J. (2018) ArXiv :1812.09150

# Contribution in our work

**Main goal :** estimation of the Wasserstein functional $W_\varepsilon(\mu, \nu)$ based on $X_1, \ldots, X_n \sim_{iid} \mu$ and assuming that $\nu$ is known

**A simple recursive estimator :**

$$\widehat{W}_n = \frac{1}{n} \sum_{k=1}^{n} h_\varepsilon(X_k, \widehat{V}_{k-1}).$$

**Rate of convergence of the expected excess risk :**

$$\widehat{R}_n = H_\varepsilon(v^*) - \mathbb{E}[\widehat{W}_n] = \frac{1}{n} \sum_{k=1}^{n} \Big( H_\varepsilon(v^*) - \mathbb{E}[H_\varepsilon(\widehat{V}_{k-1})] \Big)$$

Here, $H_\varepsilon$ is not strongly concave, but satisfies a generalized self-concordance property [1] allowing to have convergence of $\widehat{R}_n$ faster than $1/\sqrt{n}$ for $\gamma_n = \gamma/n^c$ where $\gamma > 0$ and $3/4 < c < 1$

1. Bach (2014)

# Contribution in our work [1]

## Proposition (Generalized self-concordance)

*For any $v \in \mathbb{R}^J$, we have*

$$\left\|\nabla H_\varepsilon(v) - \nabla^2 H_\varepsilon(v^*)(v - v^*)\right\| \leq \frac{1}{\varepsilon^2\sqrt{2}}\|v - v^*\|^2.$$

*Moreover, assume that $\|v - v^*\| \leq A$ for some $A > 0$. Then,*

$$\langle \nabla H_\varepsilon(v), v - v^* \rangle \leq \begin{cases} -\rho^* g\left(\frac{\sqrt{2}}{\varepsilon}A\right)\|v - v^*\|^2 & \textit{if} \quad A \leq 1, \\ -\frac{\rho^*}{A} g\left(\frac{\sqrt{2}}{\varepsilon}\right)\|v - v^*\|^2 & \textit{if} \quad A \geq 1. \end{cases}$$

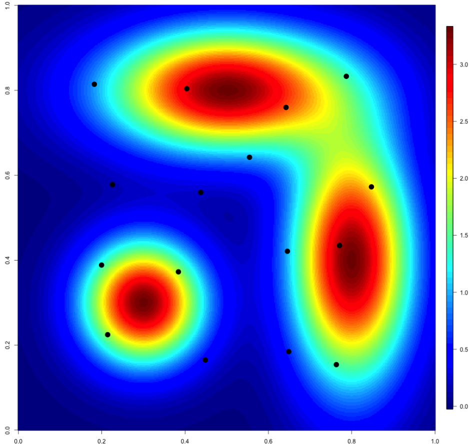
*where*

$$g(\eta) := \frac{1}{\eta}\big(1 - \exp(-\eta)\big) \geq \exp(-\eta)$$

1. Bercu, B. & Bigot, J. (2018) ArXiv :1812.09150

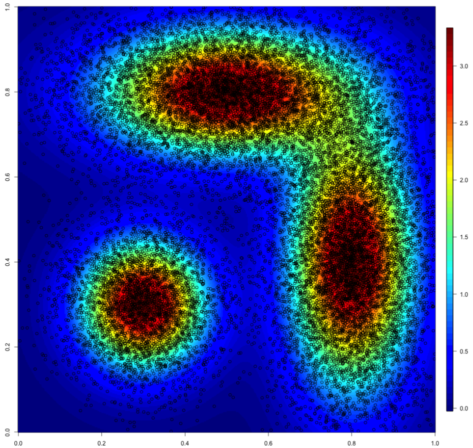
# Numerical experiments - Simulated data

Optimal transport of an absolutely continuous measure $\mu$ onto a discrete measure $\nu$ (black dots)



# Numerical experiments - Simulated data

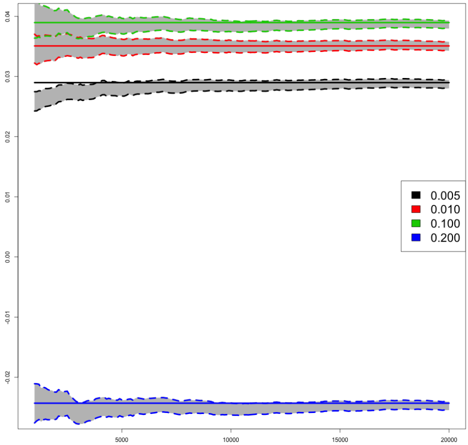
Samples $X_1, \ldots, X_N \sim_{iid} \mu$ (with $N = 20000$) and discrete measure $\nu$ (black dots)



# Numerical experiments - Simulated data

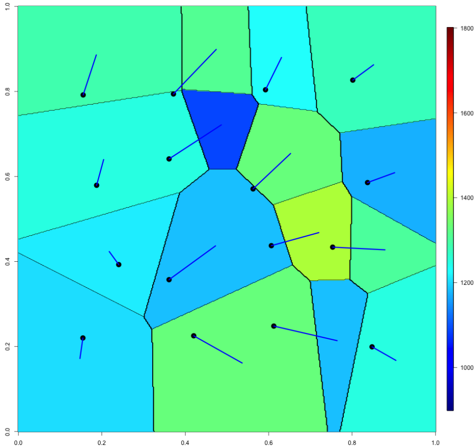
Convergence of the algorithm using the quadratic cost

$$c(x,y) = \|x-y\|_{\ell_2}^2$$



# Numerical experiments - Simulated data

Estimated optimal map $\hat{T}_{\varepsilon,N}$ for the quadratic cost $c(x,y) = \|x-y\|_{\ell_2}^2$ after $N = 20000$ iterations and $\varepsilon = 0.005$



# Numerical experiments - Laguerre cells estimation

Estimation of Laguerre cells after $n$ iterations

$$\widehat{T}_{\varepsilon,n}^{-1}(y_j) = \left\{ x \in \operatorname{supp}(\mu) \ : \ c(x,y_j) - \widehat{V}_{n,j} \leq c(x,y_k) - \widehat{V}_{n,k} \text{ for all } 1 \leq k \leq J \right\}$$

where $\widehat{V}_{n,j}$ denotes the $j$-entry of the vector $\widehat{V}_n$ considered as an estimation of a maximizer of the un-regularized semi-dual problem

$$v_0^* \in \arg\min_{v \in \mathbb{R}^J} \mathbb{E}[h_0(X,v)]$$

where $v \mapsto h_0(x,v)$ is not differentiable !

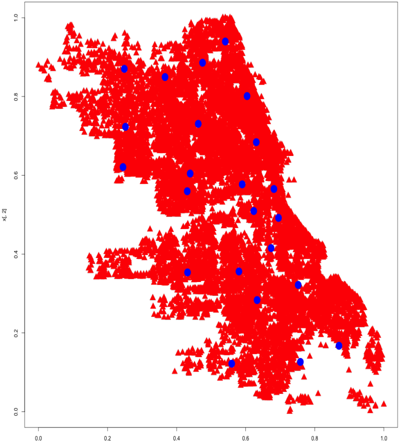
**Question (of interest ?) :** how estimating the true Laguerre cells

$$T^{-1}(y_j) = \left\{ x \in \operatorname{supp}(\mu) \ : \ c(x,y_j) - v_{j,0}^* \leq c(x,y_k) - v_{k,0}^* \text{ for all } 1 \leq k \leq J \right\}$$

bet letting $\varepsilon = \varepsilon_n \to 0$ ?

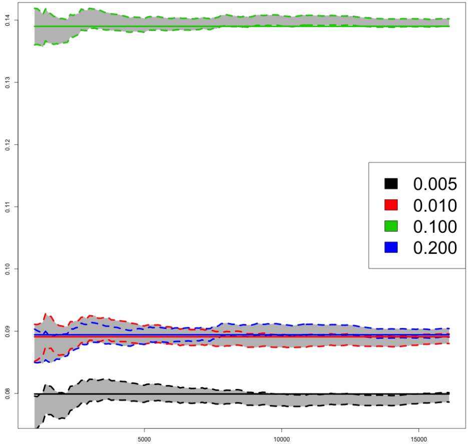
# Numerical experiments - Real data

Spatial locations $X_1, \ldots, X_N$ of reported incidents of crime in Chicago in **chronological order** (total $N = 16104$)



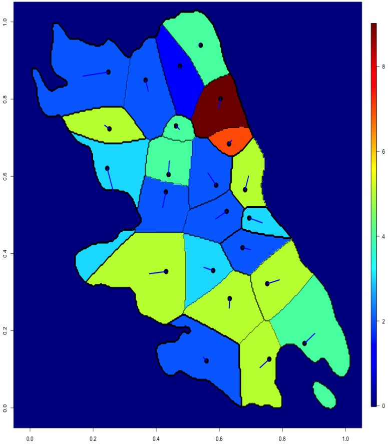
# Numerical experiments - Real data

Convergence of the algorithm using the Euclidean cost
$$c(x,y) = \|x-y\|_{\ell_2}$$



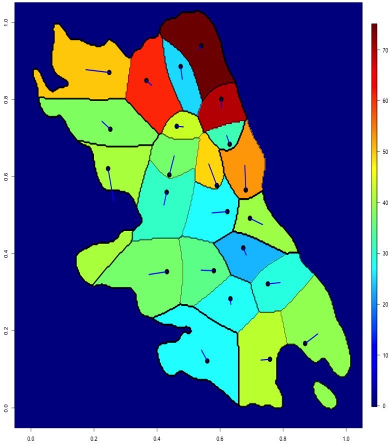
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,n}$ for the Euclidean cost $c(x,y) = \|x-y\|_{\ell_2}$
after $n = 100$ iterations and $\varepsilon = 0.005$



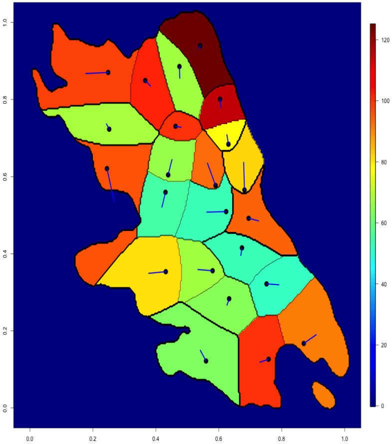
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,n}$ for the Euclidean cost $c(x,y) = \|x-y\|_{\ell_2}$ after $n = 1000$ iterations and $\varepsilon = 0.005$



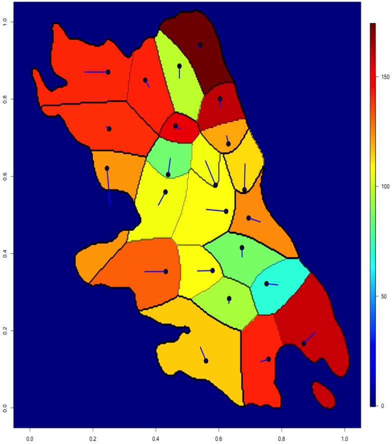
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,n}$ for the Euclidean cost $c(x,y) = \|x-y\|_{\ell_2}$ after $n = 2000$ iterations and $\varepsilon = 0.005$



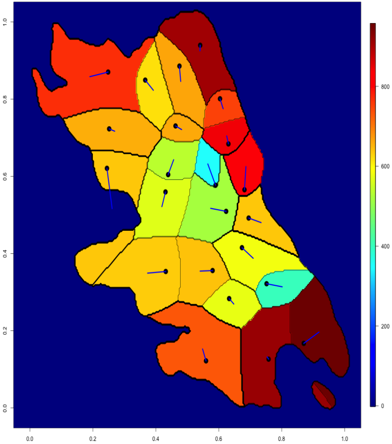
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,n}$ for the Euclidean cost $c(x,y) = \|x-y\|_{\ell_2}$ after $n = 3000$ iterations and $\varepsilon = 0.005$



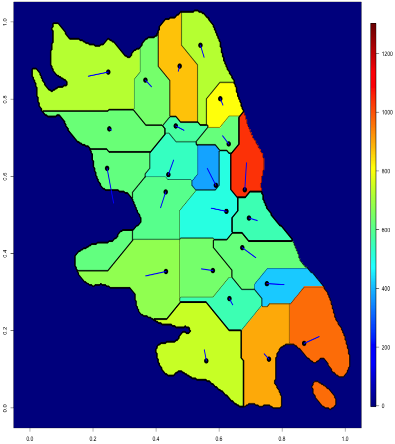
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,N}$ for the Euclidean cost $c(x,y) = \|x-y\|_{\ell_2}$ after $N = 16104$ iterations and $\varepsilon = 0.005$



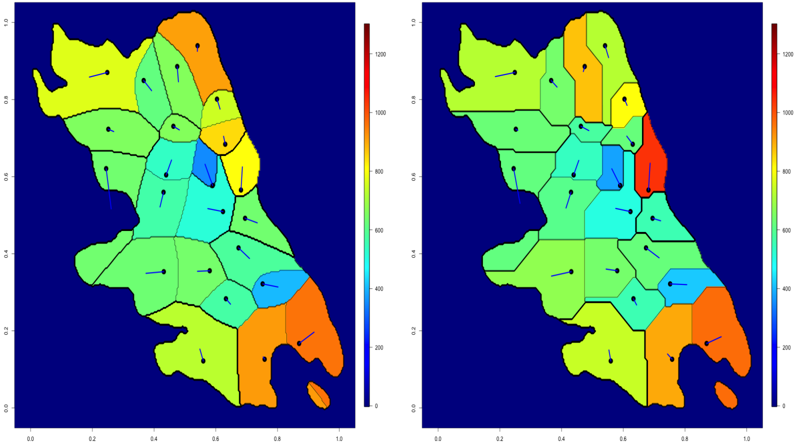
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,N}$ for the $\ell_1$ cost $c(x,y) = \|x-y\|_{\ell_1}$ after $N = 16104$ iterations and $\varepsilon = 0.005$



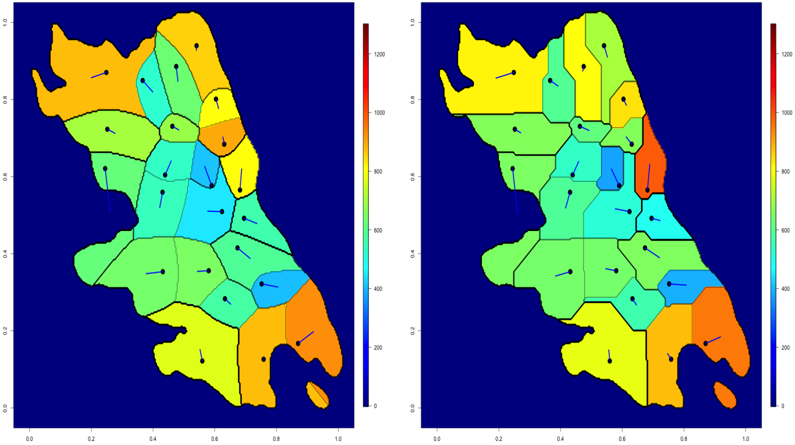
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,N}$ : Euclidean versus $\ell_1$ cost with $\varepsilon = 0.005$



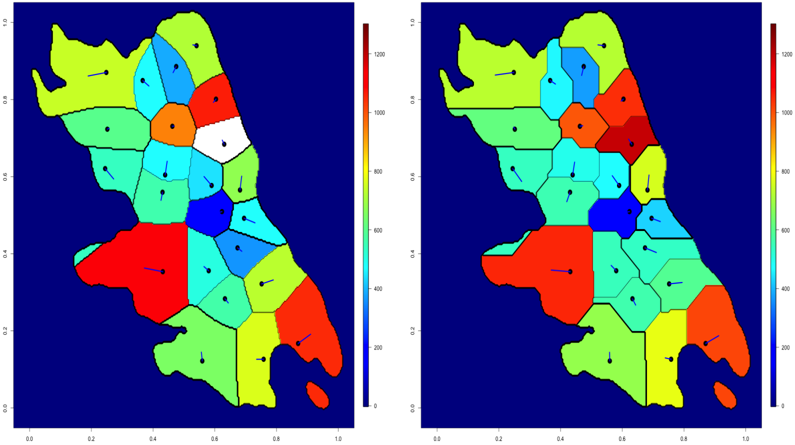
# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,N}$ : Euclidean versus $\ell_1$ cost with $\varepsilon = 0.01$



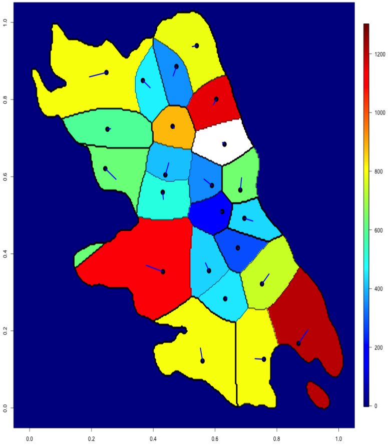
# Numerical experiments - Real data

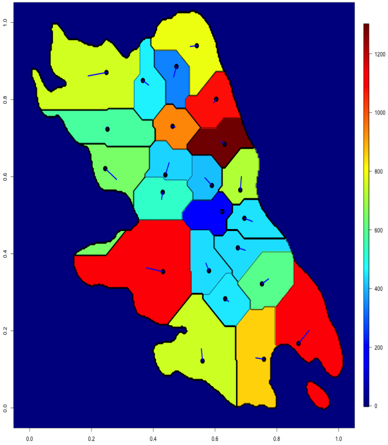
Estimated optimal map $\hat{T}_{\varepsilon,N}$ : Euclidean versus $\ell_1$ cost with $\varepsilon = 0.1$



# Numerical experiments - Real data

Estimated optimal map $\hat{T}_{\varepsilon,N}$ : Euclidean versus $\ell_1$ cost with $\varepsilon = 0.2$



# Regularized Wasserstein barycenters [1]

**Observations** of $n$ discrete measures $\tilde{\boldsymbol{\nu}}_{p_i} = \frac{1}{p_i}\sum_{j=1}^{p_i} \delta_{\mathbf{X}_{i,j}}$ for $1 \leq i \leq n$ supported on $\mathcal{X} \subset \mathbb{R}^d$.

Use of **entropically regularized** Wasserstein cost

$$\hat{\boldsymbol{\mu}}_{n,p}^{\varepsilon} = \underset{\mu \in \mathbb{P}_2(\mathcal{X})}{\text{argmin}} \ \frac{1}{n} \sum_{i=1}^{n} W_{2,\varepsilon}^2 \left(\mu, \tilde{\boldsymbol{\nu}}_{p_i}\right) \ \text{(Sinkhorn barycenter)},$$

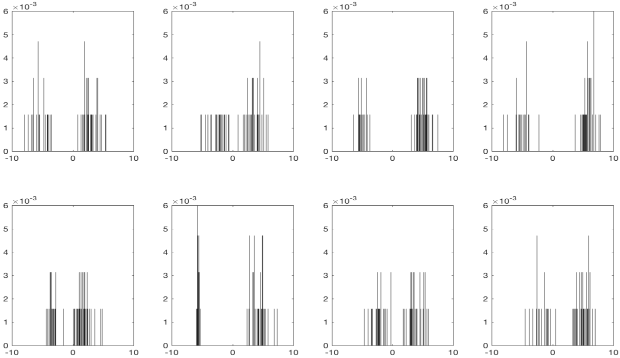
where

$$W_{2,\varepsilon}^2(\mu,\nu) = \inf_{\pi} \int_{\mathcal{X}} \int_{\mathcal{X}} |x-y|^2 \pi(x,y) dxdy - \varepsilon H(\pi),$$

where $H(\pi)$ is the entropy of the transport plan $\pi$

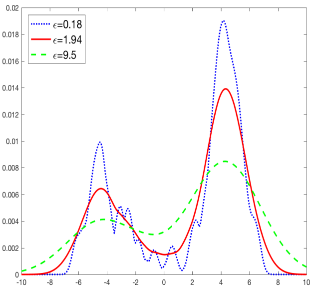
1. Bigot, J., Cazelles, E. & Papadakis, N. (2018) ArXiv :1804.08962

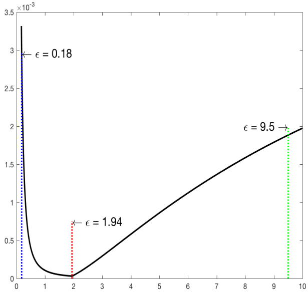
# Regularization using the Sinkhorn barycenter



A subset of $8$ histograms (out of $n = 15$) from random variables sampled from one-dimensional Gaussian mixtures distributions $\boldsymbol{\nu}_i$ (with random means and variances) and binning of the data $(\mathbf{X}_{i,j})_{1 \leq i \leq n\ ;1 \leq p}$ on a grid of size $N = 2^8$ with $p_1 = \ldots = p_n = 50$.

# Regularization using the Sinkhorn barycenter [1]




- Three Sinkhorn barycenters $\hat{\mu}^{\varepsilon}_{n,p}$ associated to the parameters $\varepsilon = 0.18, 1.94, 9.5$.
- The trade-off function $\varepsilon \mapsto \overbrace{B(\varepsilon)}^{Bias} + b\,\overbrace{V(\varepsilon)}^{Variance}$ which attains its optimum at $\hat{\varepsilon} = 1.94$ using the Goldenshluger-Lepski's principle ($L$-curve criterion)

1. Bigot, J., Cazelles, E. & Papadakis, N. (2018) ArXiv :1804.08962

# The Goldenshluger-Lepski's principle [1]

Consider a finite collection of estimators $(\hat{\boldsymbol{\mu}}_{n,p}^{\varepsilon})_{\varepsilon}$ for $\varepsilon \in \Lambda$.

The GL method consists in choosing a value $\hat{\varepsilon}$ which minimizes the bias-variance trade-off function :

$$\hat{\varepsilon} = \underset{\varepsilon \in \Lambda}{\arg\min}\ B(\varepsilon) + bV(\varepsilon)$$

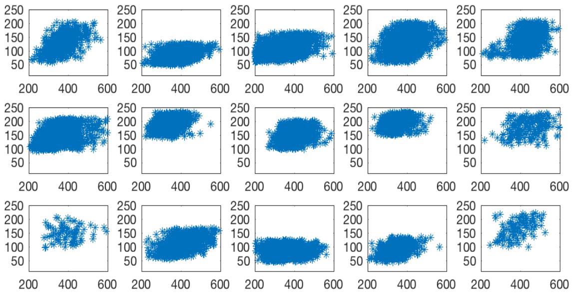
with "bias term" as

$$B(\varepsilon) = \sup_{\tilde{\varepsilon} \leq \varepsilon} \left[ |\hat{\boldsymbol{\mu}}_{n,p}^{\varepsilon} - \hat{\boldsymbol{\mu}}_{n,p}^{\tilde{\varepsilon}}|^2 - bV(\tilde{\varepsilon}) \right]_+$$

and a "variance term" $V(\varepsilon)$ chosen proportional to an upper bound on the variance of the Sinkhorn barycenter $\hat{\boldsymbol{\mu}}_{n,p}^{\varepsilon}$ (with $b > 0$ another tuning constant !)

1. e.g. for density estimation see Lacour and Massart (2016)

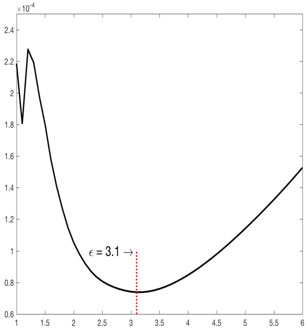
# Flow cytometry data [1]

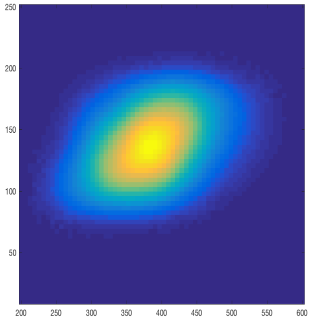



- Measurements from $n = 15$ patients restricted to a bivariate projection : FSC versus SSC cell markers.
- **Main issue :** data alignement and density estimation for cells clustering

1. Bigot, J., Cazelles, E. & Papadakis, N. (2018) ArXiv :1804.08962

# Flow cytometry data [1]



- The trade-off function $\varepsilon \mapsto \ B(\varepsilon) + bV(\varepsilon)$
- Sinkhorn barycenter associated to the parameter $\hat{\varepsilon} = 3.1$

1. Bigot, J., Cazelles, E. & Papadakis, N. (2018) ArXiv :1804.08962

# Data-driven smoothing of Laguerre cells ?

Estimated optimal map $\hat{T}_{\varepsilon,N}$ for various values of $\varepsilon$