Density Estimation and Smoothing Based on Regularized Optimal Transport

Martin Burger[1], Marzena Franek[1], Carola-Bibiane Schönlieb[2]

[1]Institute for Numerical und Applied Mathematics
University of Münster (Germany)
[2]Department for Applied Mathematics and Theoretical Physics
University of Cambridge, UK

Warwick -May, 27th 2011

UNIVERSITY OF CAMBRIDGE

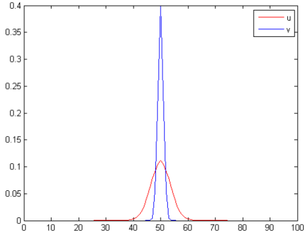
The problem

Let $\Omega \subset \mathbb{R}^d$, $d = 1, 2$ and Ω is open and bounded. We shall investigate the solution of the variational problem

$$\frac{1}{2} W_2(\nu, u\mathcal{L}^d)^2 + \alpha E(u) \to \min_u,$$

where ν is a given probability measure on Ω and u is a probability density.

E(u) is a regularizing functional . . .



$$E(u) = \int_\Omega u \log u$$

Regularizer $E(u)$

The typical regularization energy we consider is the total variation (edge preserving)

$$
\begin{aligned}
E(u) &= \int |\nabla u| \\
&:= \sup_{\mathbf{g} \in C_0^1(\Omega;\mathbb{R}^d), \|g\|_\infty \leq 1} \int_\Omega u \, \nabla \cdot \mathbf{g} \, dx.
\end{aligned}
$$

Alternatives:

- gradient squared (Dirichlet energy): $E(u) = \frac{1}{2} \int_\Omega |\nabla u|^2 \, dx$
- squared L^2-norm: $E(u) = \frac{1}{2} \int_\Omega u^2 \, dx$
- statistical functionals, log entropy: $E(u) = \int_\Omega u \ln u \, dx$
- Fisher information: $E(u) = \int_\Omega \frac{|\nabla u|^2}{u} \, dx$

The Wasserstein distance

Let $(\Omega, |\cdot|)$ constitute our metric space. The $(p-th)$ Wasserstein distance between two probability measures $\mu^1, \mu^2 \in \mathbb{P}_p(\Omega)$ is defined by

$$\boxed{W_p(\mu^1, \mu^2)^p := \min_{\Pi \in \Gamma(\mu^1, \mu^2)} \int_{\Omega \times \Omega} |x-y|^p \, d\Pi(x,y).}$$

Here $\Gamma(\mu^1, \mu^2)$ denotes the class of all transport plans $\gamma \in \mathbb{P}(\Omega^2)$ such that

$$\pi^1_\# \gamma = \mu^1, \quad \pi^2_\# \gamma = \mu^2,$$

where $\pi^i : \Omega^2 \to \Omega$, $i = 1, 2$, and $\pi^i_\# \gamma \in \mathbb{P}(\Omega)$ is the push-forward of γ through π^i.
Here: $p = 2$ quadratic Wasserstein distance $W_2(\cdot, \cdot)$ on $\mathbb{P}_2(\Omega)$.

Gradient flow formulation - JKO

Idea: Interpretation of the variational problem

$$\min_{u\mathcal{L}^d \in \mathbb{P}(X)} \mathcal{J}(u) = \frac{1}{2} W_2(\nu, u\mathcal{L}^d)^2 + \alpha E(u)$$

as one timestep of size α of the discrete solution of the gradient flow of $E(u)$ with respect to the L^2-Wasserstein-distance. It means: Solving the diffusion-equation

$$\partial_t u = \nabla \cdot (u \nabla E'(u))$$

$$u(0,x) = u_0(x) \geq 0 \quad \int_\Omega u_0 dx = 1$$

approximately using the Jordan-Kinderlehrer-Otto (JKO) scheme

$$u^{k+1} = \arg\min_u \frac{1}{2} W_2(u^k \mathcal{L}^d, u\mathcal{L}^d)^2 + (t_{k+1} - t_k) E(u), \; k \geq 0.$$

- Thin-film-equation

$$E(u) = \frac{1}{2}\int_\Omega |\nabla u|^2\, dx \Rightarrow \quad \partial_t u = div(u \nabla \Delta u)$$

- Porous Medium equation

$$E(u) = \frac{1}{2}\int_\Omega u^2 dx \Rightarrow \quad \partial_t u = \Delta u^2$$

- Heat Equation

$$E(u) = \int_\Omega u\, log(u) dx \Rightarrow \quad \partial_t u = \Delta u$$

- Derrida-Lebowitz-Speer-Spohn equation

$$E(u) = \int_\Omega u\, |\nabla log u|^2\, dx \Rightarrow \quad \partial_t u = -\Delta(u \Delta(log u))$$

- Highly nonlinear fourth order equation

$$E(u) = \int_\Omega |\nabla u|\, dx \Rightarrow \quad \partial_t u = -div\left(u \nabla div\left(\frac{\nabla u}{|\nabla u|}\right)\right)$$

Numerics based on the JKO scheme

- numerical solution of diffusion equations with schemes that respect its gradient flow structure, e.g., schemes which guarantee monotonic decrease of the corresponding energy functional,
- such schemes raised growing interest in the last years, cf., e.g., [Gosse,Toscani 06; Carrillo, Moll 09; Düring, Matthes, Milisic 10; Burger, Carrillo, Wolfram 10].

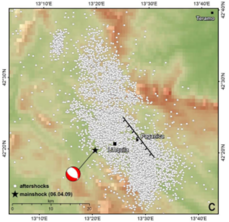
Note, however

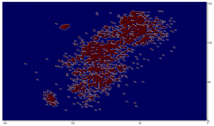
1. With our variational approach we are not aiming to compute solutions of the gradient flow! Minimizer will be solution after one timestep of size α.

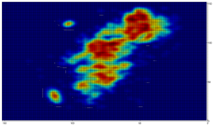
Density estimation

1. Estimation of densities from given measurements.

Example: Prediction of earthquakes (densities of intensities)







G. A. Papadopoulos, M. Charalampakis, A. Fokaefs, G. Minadakis

- Densities of intensities and locations of terrestial incidents computed from given data measured over time, see, e.g., [Ogata 1998; Egozcue, Barbat, Canas, Miquel, Banda 2006]
- Estimate crime probabilities for different districts or localities within a city [Bertozzi, Goldstein, Keegan, Mohler, Osher, Short, Smith, Wittman 2009/10]
- Wildfire predictions [Schoenberg, Chang, Keeley, Pompa, Woods, Xu 2007]

Density estimation (cont.)

State of the art techniques:

- **Maximum Penalized Likelihood Estimation (MPLE)** $\Rightarrow$ variational approaches of the form

$$\boxed{u^* \in \operatorname{argmin}_u \left[D(u,\nu) + \alpha E(u) \right],}$$

  where

  - D distance measure between density u and discrete measure ν (a sum of point densities), e.g., L^2, log-likelihood
  - $\alpha > 0$ regularisation parameter,

Density estimation (cont.)

State of the art techniques (cont.):

- **MPLE:**

$$u^* \in \operatorname{argmin}_u \left[D(u, \nu) + \alpha E(u) \right],$$

  where

  - E appropriate regularisation functional, e.g., [Good, Gaskins 1971; Eggermont, LaRiccia 2001] (model special structure of densities, e.g. discontinuities)
  - $\Rightarrow$ total variation regularization, e.g., [Koenker, Mizera 2007; Obereder, Scherzer, Kovac 2007; Mohler, Bertozzi, Goldstein, Osher 2010; Sardy, P. Tseng 2010]

- **Others:** Kernel density estimation techniques [Silverman 1982/86],taut string method, e.g., [Davies, Kovac 2004], logspline technique [Kooperberg, Stone 2002].

Connection to MPLE

Discrete version of our model motivated from a MPLE:

Let $x_1, \ldots, x_N$ be observations of N i.i.d. random variables, with unknown mean μ_i and variance σ (Gaussian error model).
Maximum a-posteriori probability estimation model derived from minimising

$$\frac{1}{2\sigma^2} \sum_i (x_i - \mu_i)^2 + \alpha E(\mu_1, \ldots, \mu_N).$$

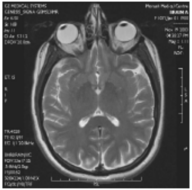
For the empirical data $\nu = \frac{1}{N} \sum_i \delta_{x_i}, \; u = \frac{1}{N} \sum_i \delta_{\mu_i}$, the squared distance term can be translated into a multiple of $W_2^2(\nu, u)$.

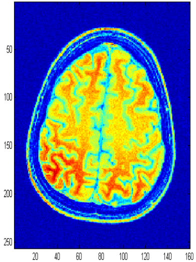
Advantage:

- Mass Conservation!
- We can work with continuous probability measures or with discrete, i.e. densities which are sums of dirac delta functions.

Smoothing of densities and cartoon-structure decomposition

2. Smoothing of noisy density images (i.e. medical images (MRI))





3. Cartoon-texture decomposition of images . . . pictures later.

Outline

1. Kantorovich formulation

Outline

1 Kantorovich formulation

2 The Benamou Brenier Ansatz

Outline

1. Kantorovich formulation
2. The Benamou Brenier Ansatz
3. Applications

Outline

1. Kantorovich formulation
2. The Benamou Brenier Ansatz
3. Applications

Regularized Kantorovich formulation

$(X, d) = (\Omega, |\cdot|)$, Kantorovich formulation with regularization:

$$J(u) = \frac{1}{2} \int_{\Omega \times \Omega} |x - y|^2 \, d\Pi(x, y) + \alpha E(u) \to \min_{\Pi, u}$$

subject to

$$\int_{A \times \Omega} d\Pi(x, y) = \int_A d\nu(y)$$

$$\int_{\Omega \times A} d\Pi(x, y) = \int_A u(x) dx$$

for $A \subset \Omega$ measurable, u probability density, ν probability measure, Π probability measure on $\Omega \times \Omega$.

Dual formulation

$$\inf_u \sup_{(\varphi,\psi)} \left(D(\varphi,\psi,u) = \int_\Omega \varphi(x)u(x)dx + \int_\Omega \psi(y)d\nu(y) + \alpha E(u) \right),$$

subject to

$$\varphi(x) + \psi(y) \leq \frac{1}{2}|x-y|^2, \tag{1}$$

$$\boxed{\inf_{\Pi\in\Gamma(\mu,\nu),u} J(u) = \inf_u \sup_{(\varphi,\psi)\in\Phi_c} D(\varphi,\psi,u)}$$

Φ_c is the set of all (φ,ψ) which satisfy (1).

Properties of the model

Existence

The functional $\mathcal{J}(u) = \frac{1}{2} W_2(\nu, u\mathcal{L}^d)^2 + \alpha E(u)$ has a minimizer $u\mathcal{L}^d \in \mathbb{P}(X)$.

Properties of the model

Existence

The functional $\mathcal{J}(u) = \frac{1}{2}W_2(\nu, u\mathcal{L}^d)^2 + \alpha E(u)$ has a minimizer $u\mathcal{L}^d \in \mathbb{P}(X)$.

Uniqueness

Let E be differentiable and strictly convex. Then there exists at most one minimizer of $\mathcal{J}(u)$.

Properties of the model

Existence

The functional $\mathcal{J}(u) = \frac{1}{2}W_2(\nu, u\mathcal{L}^d)^2 + \alpha E(u)$ has a minimizer $u\mathcal{L}^d \in \mathbb{P}(X)$.

Uniqueness

Let E be differentiable and strictly convex. Then there exists at most one minimizer of $\mathcal{J}(u)$.

Stability

Let E be differentiable and strictly convex and let $\alpha > 0$ then

$$D_E(u_1, u_2) = \alpha\langle E'(u_1) - E'(u_2), u_1 - u_2\rangle \leq W_2(\nu_1, \nu_2),$$

where D_E denotes the symmetric Bregman distance.

Convex optimization problem

$\{x_1, ..., x_n\}, \{y_1, ..., y_m\} \subset \mathbb{R}^d$, n sources, m destinations

Algorithm

$$\min_{x,u} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} p_{ij} + \alpha E(u)$$

$$\sum_{i=1}^{n} p_{ij} = v_j \quad j = 1...m, \quad \sum_{j=1}^{m} p_{ij} = u_i \quad i = 1...n$$

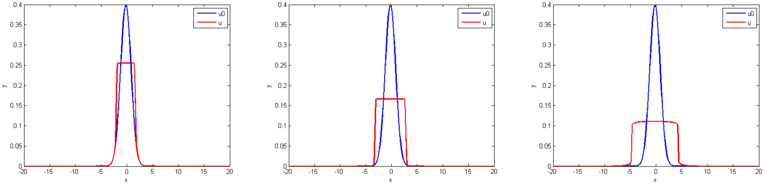
$$0 \leq p_{ij} \quad 0 \leq u_i \qquad i = 1...n, \; j = 1...m,$$

$$\sum_{ij} p_{ij} = 1, \quad \sum_{i} u_i = 1,$$

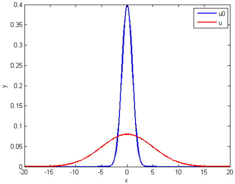
with $c_{ij} = |x_i - y_j|^2$ and $p \in \mathbb{R}^{n \times m}$. Only efficiently solvable in 1D

Numerical results

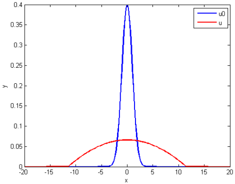
TV-Regularization $E(u) = \int_\Omega |\nabla u| \, dx \approx \int_\Omega \sqrt{|\nabla u|^2 + \epsilon^2}$



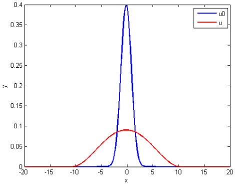
Numerical results



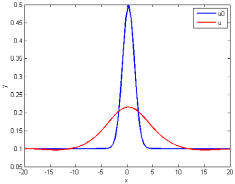
(a) Log-Entropy



(b) L^2-Reg



(c) Dirichlet-Reg



(d) Fisher-Information

Self-similar solutions

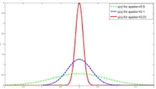
The regularised densities u we have just computed are self-similar solutions of the problem:

$$d\nu = \frac{1}{\delta^d}\, u\left(\frac{x}{\delta}\right)\, dx.$$

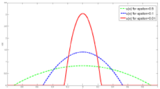
which you can explicitly compute using the Benamou-Brenier formulation:

$$\rho(x,t) = \frac{1}{(at+b)^d} u\left(\frac{x}{at+b}\right), \quad v(x,t) = \frac{ax}{at+b}, \quad \lambda(x,t) = \frac{a|x|^2}{2(at+b)}$$

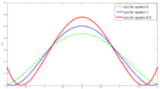
with nonnegative constants a and b such that $a+b=1$.



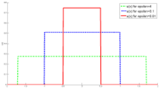
(a) Logarithmic entropy



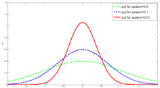
(b) Squared Lebesque regularisation



(c) Dirichlet regularisation



(d) TV regularisation



(e) Fisher Information

Outline

1 Kantorovich formulation

2 The Benamou Brenier Ansatz

3 Applications

The Benamou Brenier Ansatz

- Let $[0,T]$ be a fix time interval, $t \in [0,T], x \in \mathbb{R}^N$.
- $\rho(t,x) \geq 0$ density
- $v(t,x) \in \mathbb{R}^N$ velocity field.

$W_2^2(\rho_0, \rho_T)$ is equivalent to the infimum of

$$T \int_\Omega \int_0^T \rho(t,x)|v(t,x)|^2 dxdt$$

over all (ρ, v) subject to:

$$\partial_t \rho + \nabla \cdot (\rho v) = 0$$

$$\rho(0,\cdot) = \rho_0$$

$$\rho(T,\cdot) = \rho_T$$

Benamou-Brenier Ansatz

$$\inf_{\rho,v,u} \frac{1}{2}\int_0^1 \int_\Omega \rho(t,x)|v(t,x)|^2 dxdt + \alpha E(u)$$

subject to

$$\partial_t \rho + \nabla \cdot (\rho v) = 0, \quad \rho(t=0) = \rho_\nu, \quad \rho(t=1) = u$$

Lagrangian-function

$$\begin{aligned} L(u,\rho,v,\lambda) &= \frac{1}{2}\int_0^1 \int_\Omega \rho|v|^2 dxdt + \alpha E(u) \\ &+ \int_0^1 \int_\Omega \rho(-\partial_t \lambda + v \cdot \nabla \lambda) dxdt \\ &+ \int_\Omega \lambda(t=0)(\rho(0,\cdot) - \rho_\nu)dx + \int_\Omega \lambda(t=1)(\rho(1,\cdot) - u)dx \end{aligned}$$

Optimality conditions

$$
\begin{aligned}
L_\rho &= \frac{1}{2}v^2 - \partial_t \lambda - \nabla \lambda v = 0, && \text{adjoint equation} \\
L_v &= \rho v - \rho \nabla \lambda = 0, && \\
L_\lambda &= \partial_t \rho + \nabla \cdot (\rho v) = 0, && \text{continuity equation} \\
L_u &= \alpha E'(u) - \lambda(t=1) \ni 0, &&
\end{aligned}
$$

Uniqueness

Let E be differentiable and strictly convex and $\alpha > 0$. Then the velocity $v = \nabla \lambda$ is unique on the support of $\mu = \rho \mathcal{L}^d$.

Idea of the proof: Similar to the proof of uniqueness for mean-field games,[Lasry, Lions]

Gradient descent scheme (GD)

1. Initial: $\rho_0(t=0) = \rho_\nu$
2. Solve the continuity equation forward in time
   $\partial_t \rho^{k+1} + \nabla \cdot (\rho^{k+1} v^k) = 0 \Longrightarrow \rho^{k+1}(t=1)$
3. Solve $\alpha E'(u^{k+1}) + \lambda^{k+1}(t=1) \ni 0$ with $u^{k+1} = \rho^{k+1}(t=1)$
4. Solve the adjoint equation backward in time
   $\frac{1}{2}\left|v^k\right|^2 - \partial_t \lambda^{k+1} - \nabla \lambda^{k+1} v^k = 0$
5. Update $v^{k+1}(1+\tau) = \tau v^k + \nabla \lambda^{k+1}$.

Disadvantage: positivity constraint on u is not automatically guaranteed, i.e., instead of $\lambda + E'(u) \ni 0$ we rather have $\lambda + E'(u) + \eta \in 0$, where η is a Lagrange multiplier for the positivity constraint! Works for log-entropy though.

Dual ascent scheme (DA)

Using $v = \nabla\lambda$

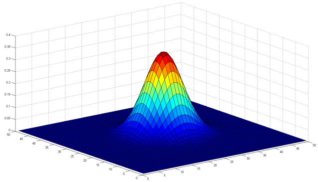
1. Initial: $\lambda(t=0) = \lambda_0$
2. Solve the adjoint equation forward in time $\frac{1}{2}\left|\nabla\lambda^{k+1}\right|^2 + \partial_t\lambda^{k+1} = 0$
3. Solve the optimization problem

$$u^{k+1} = \operatorname{argmin}_u \alpha E(u) + \int_\Omega u\lambda^{k+1}(t=1), \quad \text{and} \quad \rho^{k+1}(t=1) = u^{k+1}$$

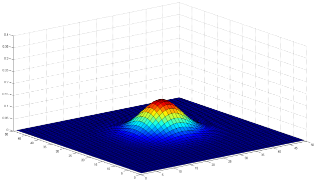
4. Solve the continuity equation backwards in time
$\partial_t\rho^{k+1} + \nabla\cdot(\rho^{k+1}\nabla\lambda^{k+1}) = 0 \Longrightarrow \rho^{k+1}(t=0)$
5. Update $\lambda_0^{k+1} = \lambda_0^k + \tau(\rho^{k+1}(t=0) - \rho_\nu)$.

Disadvantage: only works for strictly convex E.

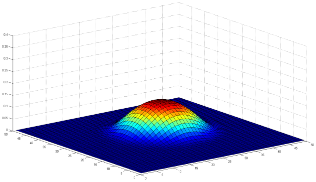
2D results



(a) initial data



(b) Log-Entropy (GD)



(c) L^2-Reg (DA)

TV-Regularisation $E(u) = \int |\nabla u|$

Augmented Lagrangian Ansatz [Glowinski, Le Tallec 89; Frick 08; Goldstein, Osher 10]

$$\min_{u,\rho,v,z} \frac{1}{2} \int_0^1 \int_\Omega \rho \, |v|^2 \, dxdt + \alpha \int_\Omega |z| \, dx,$$

subject to

$$\begin{aligned} \partial_t \rho + \nabla \cdot (\rho v) = 0, \; z = \nabla u, \\ \rho(t=0) = \rho_\nu, \; \rho(t=1) = u. \end{aligned}$$

$\Rightarrow$ saddle point problem

$$\begin{aligned} \min_{\rho,v,u,z} \max_{\xi,\lambda} L(\rho, v, u, z, \lambda) = [...] + \alpha \int_\Omega |z| \, dx \\ + \int_\Omega (z - \nabla u) \xi dx + \frac{\gamma}{2} \int_\Omega |z - \nabla u|^2 \, dx \end{aligned}$$

1. *Initialization:* $\lambda_0, v^0, \xi^0, z^0$.
2. Solve forward in time $\partial_t \lambda^{k+1} = -\frac{1}{2}\left|\nabla \lambda^{k+1}\right|^2$
3. Compute $\rho^{k+1}(t=1) = u^{k+1}$ as

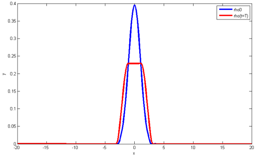
$$u^{k+1} = \Delta^{-1}\left(\frac{1}{\gamma}\lambda^{k+1}(t=1) + \frac{1}{\gamma}\nabla\cdot\xi^k + \nabla\cdot z^k\right).$$

4. Solve backwards in time $\partial_t \rho^{k+1} + \nabla\cdot(\rho^{k+1}\nabla\lambda^{k+1}) = 0$.
5. Update $\lambda_0^{k+1} = \lambda_0^k + \tau_1(\rho^{k+1}(t=0) - \rho_\nu)$
6. Shrinkage

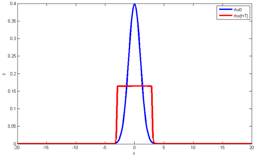
$$z^{k+1} = \begin{cases} \left(1 - \frac{\alpha}{\gamma\left|(\nabla u - \frac{\xi}{\gamma})(x,y)\right|}\right)((\nabla u - \frac{\xi}{\gamma})(x,y)), & \left|(\nabla u - \frac{\xi}{\gamma})(x,y)\right| > 1, \\ 0, & \left|(\nabla u - \frac{\xi}{\gamma})(x,y)\right| \leq 1. \end{cases}$$

7. $\xi^{k+1} = \xi^k + \tau_2(z^{k+1} - \nabla u^{k+1})$

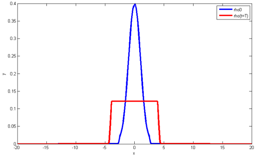
Numerical results



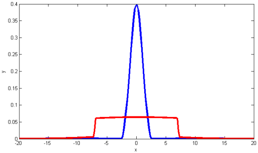
(a) $\alpha = 1$



(b) $\alpha = 5$



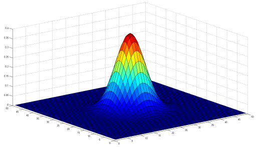
(c) $\alpha = 20$



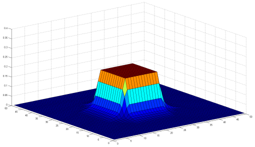
(d) $\alpha = 100$

Anisotropic TV

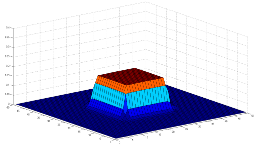
$$E(u) = \int |\nabla u| = \int |\nabla_x u| + |\nabla_y u|$$



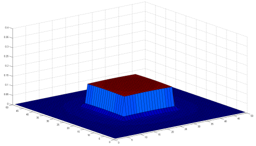
(a) $\alpha = 0$



(b) $\alpha = 10$



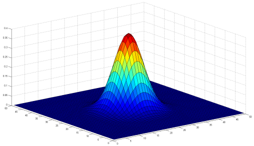
(c) $\alpha = 20$



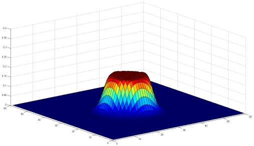
(d) $\alpha = 50$

Isotropic TV

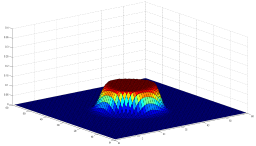
$$E(u) = \int |\nabla u| = \int \sqrt{(\nabla_x u)^2 + (\nabla_y u)^2}$$



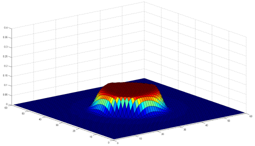
(a) $\alpha = 0$



(b) $\alpha = 10$



(c) $\alpha = 20$



(d) $\alpha = 50$

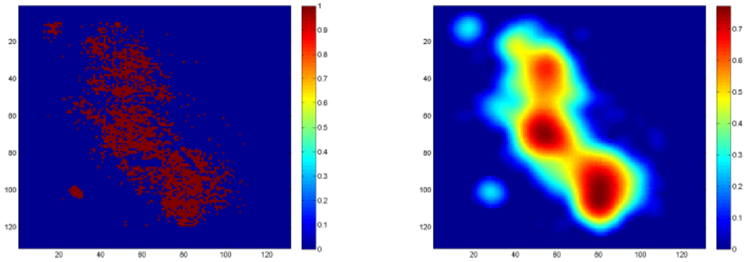
Outline

1 Kantorovich formulation

2 The Benamou Brenier Ansatz

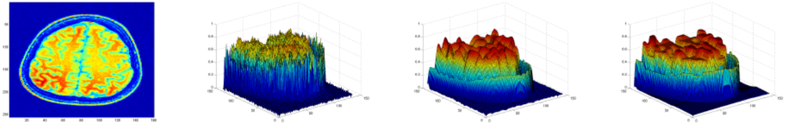
3 Applications

Density estimation with Dirichlet-Wasserstein

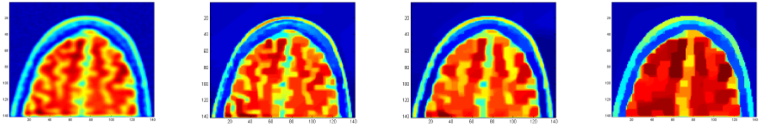


(a) Initial data (b) Dirichlet regularisation

MRI density smoothing with TV-Wasserstein



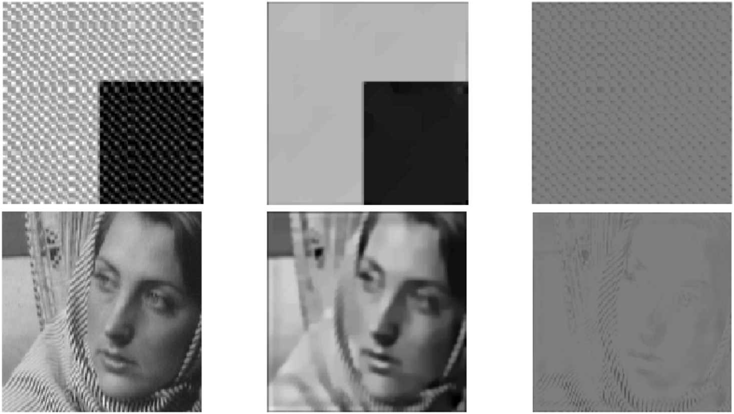
(a) MRI measurements (b) Initial data - 150×150 detail (c) Dirichlet reg. (d) TV - $\epsilon = 0.2$



(e) Dirichlet reg. (f) TV - $\epsilon = 0.05$ (g) TV - $\epsilon = 0.2$ (h) TV - $\epsilon = 1$

Thanks to Martin Uecker from the MPI for biophysical chemistry in Göttingen for providing us with the MRI data.

Image decomposition with TV-Wasserstein



Conclusion

- Regularized optimal transport: Minimize Wasserstein distance plus regularizing functional (e.g., for denisity estimation)
- Existence & uniqueness results
- Numerical scheme based on Benamou-Brenier
- Density estimation and smoothing for real-world problems.

Conclusion

- Regularized optimal transport: Minimize Wasserstein distance plus regularizing functional (e.g., for denisity estimation)
- Existence & uniqueness results
- Numerical scheme based on Benamou-Brenier
- Density estimation and smoothing for real-world problems.

Thank you for your attention!

Email: cbs31@cam.ac.uk

Reference: M. Burger, M. Franek, C.-B. Schönlieb, *Density estimation and smoothing based on regularised optimal transport*, submitted 2011.