Discrete transport problems
and the concavity of entropy

Oliver Johnson and Erwan Hillion
University of Bristol and University of Luxembourg

Bristol Probability and Statistics Seminar, March 2014

Funded by EPSRC *Information Geometry of Graphs* EP/I009450/1
Paper arXiv:1303.3381
Motivating Problem I

- Suppose we have a pile of soil we need to move somewhere (say along $\mathbb{R}$, or along $\mathbb{Z}$).
- Each spadeful moved from point $x$ to point $y$ costs us something.
- Fix cost function e.g. $c(t) = |t|^p$ for $p \geq 1$.
- Moving one spadeful from $x$ to $y$ costs $c(y - x) = |y - x|^p$.
- Can we find a moving strategy that minimises the total cost ...?
... Yes We Can!

- Source and destination piles need to have same size.
- Suppose piles have the same shape.
- Intuitive solution: just translate (move everything same distance).
More general case

- What if piles not the same shape?
- For many cost functions $c$, intuitive ‘non-crossing principle’.
Non-crossing principle

Suppose spadefuls of soil at $w$ and $x$ to move to $y$ and $z$.

Take cost $c(t) = t^2$ for simplicity.

Strategy 1: $w \rightarrow z$, $x \rightarrow y$. Cost $c_1 = (z - w)^2 + (y - x)^2$.

Strategy 2: $w \rightarrow y$, $x \rightarrow z$. Cost $c_2 = (y - w)^2 + (z - x)^2$.

$c_1 - c_2 = 2(x - w)(z - y) \geq 0$.

Prefer Strategy 2: not to let soil cross over.

Similar argument for any convex cost function.
Transport of probability measures

- Can rephrase problem more mathematically.
- Transport probability density function (or mass function) $f_0$ to $f_1$.
- Equivalently think in terms of distribution functions $F_0$ and $F_1$.
- Write $\Gamma(F_0, F_1)$ for the set of joint probability distributions with marginals $F_0$ and $F_1$ (couplings).
- Joint density $f(x, y)$ codes the amount of mass to be moved from $x$ to $y$ for particular strategy.
Transport of probability measures on \( \{0, 1\} \)

**Example**

- Consider marginals \( f_0 = (3/4, 1/4) \) and \( f_1 = (1/4, 3/4) \).
- Could define \( f(x, y) \) as follows:

<table>
<thead>
<tr>
<th>( f(x, y) )</th>
<th>( y = 0 )</th>
<th>( y = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( x = 0 )</td>
<td>1/4</td>
<td>1/2</td>
</tr>
<tr>
<td>( x = 1 )</td>
<td>0</td>
<td>1/4</td>
</tr>
</tbody>
</table>

\( f_0(0) = 3/4 \quad f_0(1) = 1/4 \)
\( f_1(0) = 1/4 \quad f_1(1) = 3/4 \)

- Cost of strategy \( f \) is

\[
\sum_{x,y} f(x, y) c(y - x) = \sum_{x,y} f(x, y) |y - x|^p.
\]
Distance between probability measures

- This gives us a way to measure how similar $F_0$ and $F_1$ are . . .
- . . . measure cost to move one distribution to the other . . .
- . . . under optimal strategy.
Distance between probability measures

**Definition**

Given $F_0$ and $F_1$ and cost function $c(t) = |t|^p$, write

$$W_p(F_0, F_1) = \left( \inf_{F \in \Gamma(F_0, F_1)} \int |y - x|^p dF(x, y) \right)^{1/p}.$$ 

- Using non-crossing principle, optimal strategy gives

  $$W_p(F_0, F_1) = \left( \int_0^1 |F_0^{-1}(t) - F_1^{-1}(t)|^p dt \right)^{1/p}.$$ 

- This is the Wasserstein distance . . . (or Vasershtein) . . . (or earth mover’s) . . . (or Mallows) . . . (or Kantorovich) . . . (or Kantorovich-Rubinstein) . . . (or Monge-Kantorovich) . . . (or Tanaka) . . . (or transport) . . . (or transportation) . . . SEO hell!
Motivating problem II

- Suppose want to run from $x = 0$ to $x = D$ in $T$ units of time.
- Suppose to maintain a speed of $v$ costs us $v^2$ in energy.
- What is correct speed to run to minimise total energy use?
- Represent trajectory in terms of a function $x(t)$, with $x(0) = 0$ and $x(T) = D$.
- Wish to minimise $\int_0^T x'(t)^2 dt$. 

Oliver Johnson and Erwan Hillion

University of Bristol and University of Luxembourg
Constant speed paths

- Obvious strategy: \( x(t) = \frac{tD}{T} \).
- Gives \( \int_0^T x'(t)^2 \, dt = T(D/T)^2 = D^2/T \).
- Now by Cauchy-Schwarz:
  \[
  \left( \int_0^T 1 \, dt \right) \left( \int_0^T x'(t)^2 \, dt \right) \geq \left( \int_0^T x'(t) \, dt \right)^2 = D^2
  \]
- That is \( T \int_0^T x'(t)^2 \, dt \geq D^2 \).
- Obvious strategy (constant speed path) is optimal.
What does this tell us about the Wasserstein distance?

- We saw how to move probability density \( f_0 \) to \( f_1 \) on \( \mathbb{R} \).
- Can think of this as taking 1 unit of time.
- Now suppose that we interrupt the process at time \( t \).
- Where would we have got to?
- Can use ideas from fluid dynamics.
- Benamou–Brenier proved variational characterization of \( W_2 \).
- Works for \( \mathbb{R} \), \( \mathbb{R}^d \), Riemannian manifolds . . . but not e.g. \( \mathbb{Z} \).
Benamou–Brenier formula

- Given distribution functions \( F_0 \) and \( F_1 \), write \( P_{\mathbb{R}}(F_0, F_1) \) for the set of densities \( f_t(x) \) such that \( F_0(x) = \int_{-\infty}^{x} f_0(y) \, dy \) and \( F_1(x) = \int_{-\infty}^{x} f_1(y) \, dy \).
- Given a sequence of densities, define velocity field \( v_t(x) \) by

\[
\frac{\partial}{\partial t} f_t(x) = -\frac{\partial}{\partial x} (v_t(x)f_t(x)).
\]

Theorem (Benamou–Brenier)

The quadratic Wasserstein distance on \( \mathbb{R} \) is given by

\[
W_2(F_0, F_1) = \left( \inf_{f_t \in P_{\mathbb{R}}(F_0, F_1)} \int_0^1 \left( \int_{-\infty}^{\infty} f_t(y) v_t(y)^2 \, dy \right) \, dt \right)^{1/2}.
\]
Benamou–Brenier geodesics

- If $f_t \in \mathcal{P}_\mathbb{R}(F_0, F_1)$ achieves infimum in Benamou–Brenier, call it a geodesic.
- Geodesics have nice properties.

Theorem

Geodesics satisfy fixed speed property:

$$W_2(F_s, F_t) = |t - s| W_2(F_0, F_1), \quad \text{for all } s \text{ and } t.$$ 

- Say that $W_2$ induces a length space.
- Fits with idea that geodesics are straight lines.
Entropy

Definition
Recall we measure ‘randomness’ of probability density $f$ by entropy

$$H(f) = - \int_{\mathbb{R}} f(x) \log f(x) dx.$$ 

- Interested in how entropy varies along paths $f_t$.
- In particular, what is behaviour along geodesics?
Behaviour of entropy along paths

Definition
Given a path $f_t(x)$, introduce functions $g_t$ and $h_t$ such that

$$\frac{\partial f_t(x)}{\partial t} = -\frac{\partial g_t(x)}{\partial x}, \quad \frac{\partial^2 f_t(x)}{\partial t^2} = \frac{\partial^2 h_t(x)}{\partial x^2}.$$ 

Theorem
Writing $H(t) = H(f_t)$ for the entropy along the path, under integrability conditions:

$$H''(t) = -\int_{\mathbb{R}} \left( h_t(x) - \frac{g_t(x)^2}{f_t(x)} \right) \frac{\partial^2}{\partial x^2} \left( \log f_t(x) \right) dx$$

$$- \int_{\mathbb{R}} f_t(x) \left( \frac{\partial}{\partial x} \left( \frac{g_t(x)}{f_t(x)} \right) \right)^2 dx.$$
Behaviour of entropy along paths

Proof.

\[
H'(t) = - \int_{\mathbb{R}} \frac{\partial f_t(x)}{\partial t} \log f_t(x) \, dx
\]

\[
H''(t) = - \int_{\mathbb{R}} \frac{\partial^2 f_t(x)}{\partial t^2} \log f_t(x) \, dx - \int_{\mathbb{R}} \frac{1}{f_t(x)} \left( \frac{\partial f_t(x)}{\partial t} \right)^2 \, dx
\]

\[
= - \int_{\mathbb{R}} \frac{\partial^2 h_t(x)}{\partial x^2} \log f_t(x) \, dx - \int_{\mathbb{R}} \frac{1}{f_t(x)} \left( \frac{\partial g_t(x)}{\partial x} \right)^2 \, dx.
\]

Integration by parts deals with these terms.

Key is an explicit expression for \( \frac{\partial^2}{\partial x^2} \log f_t(x) \).
Behaviour of entropy along geodesics

- Along BB geodesics turns out \( g_t(x) = v_t(x)f_t(x) \) and \( h_t(x) = v_t(x)^2f_t(x) \).
- In above theorem

\[
H''(t) = - \int_{\mathbb{R}} f_t(x) \left( \frac{\partial v_t(x)}{\partial x} \right)^2 \, dx \leq 0.
\]

- This concavity used in information geometry.
- Properties of \( W_2 \) are key.
- Special case of Sturm–Lott–Villani theory. For example:

**Theorem**

*For a Riemannian manifold \((M, d)\) concavity of entropy along every geodesic is equivalent to positivity of the Ricci curvature tensor.*
Discrete random variables

- Situation less clear for random variables supported on discrete sets.
- Will consider random variables supported on $\mathbb{Z}$ ...
- ... or in fact $\{0, 1, \ldots, n\}$.
For discrete problems, $W_2$ is not a length space

Example

- Consider marginals $f_0 = (3/4, 1/4)$ and $f_1 = (1/4, 3/4)$
- Obvious (and optimal) strategy $f_t = (3/4 - t/2, 1/4 + t/2)$.
- Could define $f_t(x, y)$ as follows:

<table>
<thead>
<tr>
<th>$f_t(x, y)$</th>
<th>$y = 0$</th>
<th>$y = 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x = 0$</td>
<td>$3/4 - t/2$</td>
<td>$t/2$</td>
</tr>
<tr>
<td>$x = 1$</td>
<td>$0$</td>
<td>$1/4$</td>
</tr>
</tbody>
</table>

  $f_t(0) = 3/4 - t/2$  $f_t(1) = 1/4 + t/2$

- Cost of $f_t$ is $W_2^2(F_0, F_t) = \sum_{x,y} f_t(x, y)|y - x|^p = t/2$.
- Hence $W_2(F_0, F_t) = \sqrt{t} W_2(F_0, F_1)$ – not a length space.
Consider $n$ independent Bernoulli random variables, with parameters $\mathbf{p} = (p_1, \ldots, p_n)$.

Their sum has mass function $f_\mathbf{p}(k)$ for $k = 0, 1, \ldots, n$.

Consider the entropy of $f_\mathbf{p}$, defined by

$$H(\mathbf{p}) := -\sum_{k=0}^{n} f_\mathbf{p}(k) \log f_\mathbf{p}(k).$$

Conjecture (Shepp–Olkin (1981))

$H(\mathbf{p})$ is a concave function of $\mathbf{p}$.

Sufficient to consider concavity for affine $t$, i.e. take

$$p_i(t) = p_i(0)(1 - t) + p_i(1)t.$$
Known cases

- Folklore: \( n = 1 \).
- Shepp–Olkin (1981): \( n = 2 \), \( n = 3 \) (claim with no proof, in paper).
- Shepp–Olkin (1981): for all \( i \), \( p_i(t) = t \) (binomial case).
- Yu–Johnson (2009): for all \( i \), either \( p_i(0) = 0 \) or \( p_i(1) = 0 \).
- Hillion (2012): for all \( i \), either \( p_i(t) = t \) or \( p_i(t) \) constant (binomial translation case).
Motivating example: binomial case

Example

- Write spatial derivative $\nabla_1 f(k) = f(k) - f(k - 1)$.
- For $0 \leq p < q \leq 1$, define $p(t) = p(1-t) + qt$.
- Write $\text{Bin}_{n,p}(k) := \binom{n}{k} p^k (1-p)^{n-k}$.
- Write $f_t(k) = \text{Bin}_{n,p(t)}(k)$.
- Simple calculation (e.g. Mateev, Shepp–Olkin) shows:

$$\frac{\partial f_t(k)}{\partial t} = -\nabla_1 \left( n(q-p) \text{Bin}_{n-1,p(t)}(k) \right).$$
Motivating example: binomial case (cont.)

Example

- We rewrite this using an idea of Yu:

\[
\text{Bin}_{n-1,p}(k) = \frac{(k + 1)}{n} \text{Bin}_{n,p}(k + 1) + \left(1 - \frac{k}{n}\right) \text{Bin}_{n,p}(k).
\]

- Suggests we introduce mixtures of mass functions:

\[
\frac{\partial f_t(k)}{\partial t} = -\nabla_1 \left( v g^{(\alpha)}_t(k) \right),
\]

for

\[
g^{(\alpha)}_t(k) = \alpha_t(k + 1)f_t(k + 1) + (1 - \alpha_t(k))f_t(k)
\]

- Here \( \alpha_t(k) = k/n \) for all \( k \) and \( t \) and \( v = n(q - p) \).

- Remember continuous equation \( \frac{\partial}{\partial t} f_t(x) = -\frac{\partial}{\partial x} (v_t(x)f_t(x)) \).
Discrete Benamou–Brenier formula

Definition

- Write $\mathcal{P}_Z(f_0, f_1)$ for the set of probability mass functions $f_t(k)$, given end constraints $f_t(k)|_{t=0} = f_0(k)$ and $f_t(k)|_{t=1} = f_1(k)$.
- Write $\mathcal{A}$ for the set of $\alpha(k)$ with $\alpha_t(0) \equiv 0$, $\alpha_t(n) \equiv 1$ and with $0 \leq \alpha_t(k) \leq 1$ for all $k$. 
Discrete Benamou–Brenier formula

Definition

For $f_t(k) \in P_\mathbb{Z}(f_0, f_1)$ and $\alpha \in A$, define probability mass function $g_t^{(\alpha)}(k)$, velocity field $v_{\alpha, t}(k)$ and distance $V_n$ by

$$g_t^{(\alpha)}(k) = \alpha_t(k + 1)f_t(k + 1) + (1 - \alpha_t(k))f_t(k)$$

$$\frac{\partial f_t}{\partial t}(k) = -\nabla_1 \left( v_{\alpha, t}(k)g_t^{(\alpha)}(k) \right)$$

$$V_n(f_0, f_1) = \left( \inf_{f_t \in P_\mathbb{Z}(f_0, f_1), \alpha_t(k) \in A} \int_0^1 \left( \sum_{k=0}^{n-1} g_t^{(\alpha)}(k)v_{\alpha, t}(k)^2 \right) dt \right)^{1/2}$$

Refer to any path achieving the infimum as a geodesic.
Discrete Benamou–Brenier formula

Definition

- Example: binomial path is geodesic with $v_{\alpha,t}(k) \equiv n(q - p)$.
- Call path with $v_{\alpha,t}(k)$ fixed in $k$ and $t$ a constant speed path.

Proposition

- $V_n$ is a metric for probability measures on $\{0, \ldots, n\}$.
- $V_n$ defines a length space: for any geodesic $f$, distance $V_n(f_s, f_t) = |t - s| V_n(f_0, f_1)$.
- If there exists a constant speed path then
  - $f_0$ and $f_1$ are stochastically ordered.
  - Wasserstein distance $W_1$ and $V_n$ coincide.
Framework for concavity of entropy

- Want conditions under which entropy is concave.
- Give conditions in terms of $\alpha_t(k)$ to generalize binomial case.
- Recall that in that case, $\alpha_t(k) \equiv k/n$. 
\textit{k-monotonicity condition}

\textbf{Condition (k-MON)}

\textit{Given }$t$\textit{, we say that the }$\alpha_t(k)$\textit{ are }$k$\textit{-monotone at }$t$\textit{ if}

$$\alpha_t(k) \leq \alpha_t(k + 1) \quad \text{for all } k = 0, \ldots, n - 1.$$
**t-monotonicity condition**

**Condition (t-MON)**

Given $t$, we say that the $\alpha_t(k)$ are $t$-monotone at $t$ if

$$\frac{\partial \alpha_t(k)}{\partial t} \geq 0 \quad \text{for all } k = 0, \ldots, n.$$ 

- Given a constant speed path
  $$\frac{\partial f_t(k)}{\partial t} = -v \nabla_1 \left( g_t^{(\alpha)}(k) \right),$$
  introduce $h(k)$ such that
  $$\frac{\partial^2 f_t(k)}{\partial t^2} = v^2 \nabla_1^2 (h(k)).$$

- $t$-MON condition provides an upper bound on $h(k)$. 

Oliver Johnson and Erwan Hillion
University of Bristol and University of Luxembourg

Discrete transport problems and the concavity of entropy
GLC condition

Condition (GLC)

We say $f_t(k)$ is $\alpha$-generalized log-concave at $t$, if for all $k = 0, \ldots, n - 2$,

$$GLC(\alpha_t)(k) := \alpha_t(k + 1)(1 - \alpha_t(k + 1))f_t(k + 1)^2$$

$$-\alpha_t(k + 2)(1 - \alpha_t(k))f_t(k)f_t(k + 2)$$

$$\geq 0.$$
Theorem (Hillion–Johnson 2014)

Consider constant speed path $f_t(k)$ and associated optimal $\alpha(t)$. If Conditions k-MON, t-MON and GLC hold at given $t = t^*$, the entropy $H(f_t)$ is concave in $t$ at $t = t^*$. 
Proof

▶ Dealing with logarithm remains key – but harder.
▶ \(k\text{-MON and GLC together imply that}\)

\[
\frac{f_t(k)g_t(k + 1)}{f_t(k + 1)g_t(k)} \leq 1 \quad \text{and} \quad \frac{f_t(k + 2)g_t(k)}{f_t(k + 1)g_t(k + 1)} \leq 1.
\]

▶ Also \(- \log \nu \leq \theta(\nu) = 1/(2\nu) - \nu/2, \text{ for } \nu \leq 1.\)
▶ Hence

\[
- \log \left(\frac{f_t(k)f_t(k + 2)}{f_t(k + 1)^2}\right) = - \log \left(\frac{f_t(k)g_t(k + 1)}{f_t(k + 1)g_t(k)}\right) - \log \left(\frac{f_t(k + 2)g_t(k)}{f_t(k + 1)g_t(k + 1)}\right) \leq \theta \left(\frac{f_t(k)g_t(k + 1)}{f_t(k + 1)g_t(k)}\right) + \theta \left(\frac{f_t(k + 2)g_t(k)}{f_t(k + 1)g_t(k + 1)}\right)
\]
Proof (cont.)

\[ H''(t) = \sum_{k=0}^{n} \frac{\partial^2 f_t(k)}{\partial t^2} \log f_t(k) - \sum_{k=0}^{n} \frac{1}{f_t(k)} \left( \frac{\partial f_t(k)}{\partial t} \right)^2 \]

\[ = -\sum_{k=0}^{n} v^2 \nabla_1^2 (h_t(k)) \log f_t(k) - \sum_{k=0}^{n} \frac{(\nabla_1(vg_t(k)))^2}{f_t(k)} \]

\[ = v^2 \sum_{k=0}^{n} h_t(k) \left( -\log \left( \frac{f_t(k)f_t(k+2)}{f_t(k+1)^2} \right) \right) - \sum_{k=0}^{n} \frac{(\nabla_1(vg_t(k)))^2}{f_t(k)} \]

\[ \leq v^2 \sum_{k=0}^{n} h_t(k) \left( \theta \left( \frac{f_t(k)g_t(k+1)}{f_t(k+1)g_t(k)} \right) + \theta \left( \frac{f_t(k+2)g_{t}(k+1)}{f_t(k+1)g_t(k+1)} \right) \right) \]

\[ - \sum_{k=0}^{n} \frac{(\nabla_1(vg_t(k)))^2}{f_t(k)} \]
Proof (cont.)

... and then, as if by magic, this becomes minus a perfect square!!

Details best left to Mathematica . . .

\( H''(t) \) becomes \( \leq -v^2 \) times . . .

\[
\sum_{k=0}^{n-2} \frac{f_t(k)f_t(k+1)f_t(k+2)}{2g_t(k)g_t(k+1)} \left( \frac{g_t(k)^2}{f_t(k)f_t(k+1)} - \frac{g_t(k+1)^2}{f_t(k+1)f_t(k+2)} \right)^2
\]

Would like to know how to interpret this cf (above)

\[
H''(t) = -\int f_t(x) \left( \frac{\partial v_t(x)}{\partial x} \right)^2 dx \leq 0.
\]
Relating this to Shepp–Olkin

Proposition

For Shepp–Olkin interpolations, if all $p'_i$ have the same sign (‘monotone case’):

- We have a constant speed path
- $k$-MON condition holds.
- GLC condition holds.
- However, $t$-MON condition fails for some Shepp–Olkin paths.
- Entropy remains concave if replace by $t$-MON by weaker ‘Condition 4’.
- Condition 4 holds for Shepp–Olkin paths.
Main result of our paper

**Theorem (Hillion–Johnson 2014)**

*If all $p'_i$ have the same sign, $H(p)$ is a concave function of $p$.***

- Call this monotone Shepp–Olkin theorem.
- General case remains open (not constant speed path).