GRETSI 2013
Brest, Bretagne, France
6 September 2013

Transportation distance, Shannon information, and source coding

Robert M. Gray
Alcatel-Lucent Technologies Professor of Engineering, Emeritus
Stanford University
Research Professor, Boston University

Suppose random objects $X \in \underbrace{A_X}_{\text{alphabet}}$, $Y \in A_Y$ have distributions μ_X, μ_Y

What is a *useful* notion of *"distance"* $d(\mu_X, \mu_Y)$?

For example . . .

- How *different* are two coin flips? single flip, n iid flips, iid forever

random variables, vectors, *processes*

- How different are 2 Gaussian random
  - variables
  - vectors ?
  - *processes*

How good is a discrete approximation to a Gaussian process?

*What is the **best** such approximation within some class $\mathcal{P}$?*

$$d(\mu_X, \mathcal{P}) = \inf_{\nu \in \mathcal{P}} d(\mu_X, \nu)$$

There are *hundreds* of notions of "distance" between measures, including probability measures. E.g.,

Kullback-Leibler divergence/relative entropy, L_p, variation, total variation, Kolmogorov, Hellinger, Prohorov, Csiszár, f-divergences, Ali-Silvey, . . .

This talk emphasizes one of them:

the transportation distance (**Monge, Kantorovich**, Wasserstein/Vasershtein, **Ornstein**, d-bar, Mallows, earth mover's, etc.) *from the viewpoint of information theory and signal processing*

– especially quantization

Why are distances on distributions useful?

- Quantify goodness of approximation of a simple probabilistic model to a complicated one.
- Describe complex families of random processes as closures of simple families.
- Use with minimum distance/nearest neighbor rules —

  source coding/quantization, detection, classification, recognition
- Provide a geometry of random processes which can be useful in stating and interpreting results in coding and signal processing.
- Quantify *continuity* of attributes (e.g. entropy/information, distortion vs. rate performance) of random processes with respect to the distance.
- Bounding performance difference between empirical distributions and true distribution in statistical learning [Pollard (1982)]

Transportation Distance

Suppose that $d(x, y) \geq 0$ assigns a *cost* to each pair $x \in A_X, y \in A_Y$ for transporting/changing/mapping/transforming/coding x into y

Information theory/signal processing point of view: d is a *distortion measure* in the Shannon sense e.g.,

- *Hamming distance* $d(x, y) = d_H(x, y) = \begin{cases} 0 & x = y \\ 1 & \text{otherwise} \end{cases}$ (metric)
- *Average Hamming distance* $x = (x_0, \cdots, x_{N-1}), y = (y_0, \cdots, y_{N-1})$
$d(x, y) = N^{-1} \sum_{n=0}^{N-1} d_H(x_i, y_i)$ (metric)
- *Squared error* $A_X = A_Y = \mathbb{R}$, $d(x, y) = |x - y|^2$ (metric^2)

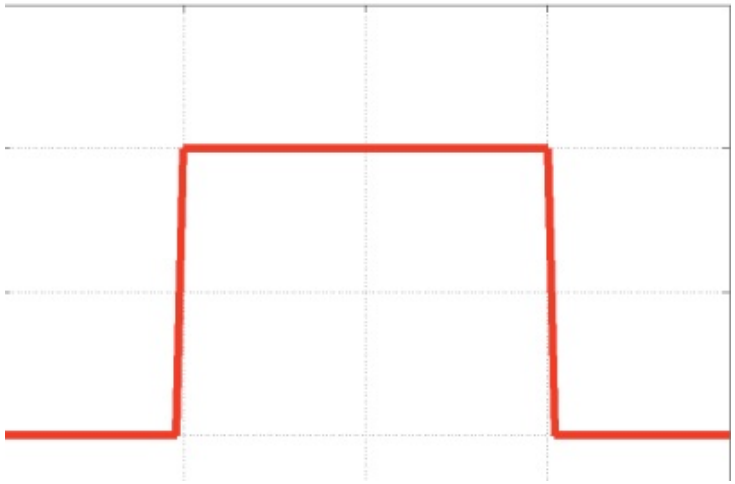
$A_X = A_Y = \mathbb{R}^N$, $d(x, y) = \|x - y\|_2^2 = \sum_{n=0}^{N-1} |x_i - y_i|^2$ (metric^2)

Monge's problem (1781): $d_{\text{Monge}}(\mu_X, \mu_Y) \equiv \inf_{f:\mu_{f(X)}=\mu_Y} E_{\mu_X}\left[d(X, f(X))\right]$

μ_X, μ_Y fixed. *Mapping* f $\quad X \longrightarrow \boxed{f} \longrightarrow Y = f(X) \qquad \mu_{f(X)} = \mu_Y$

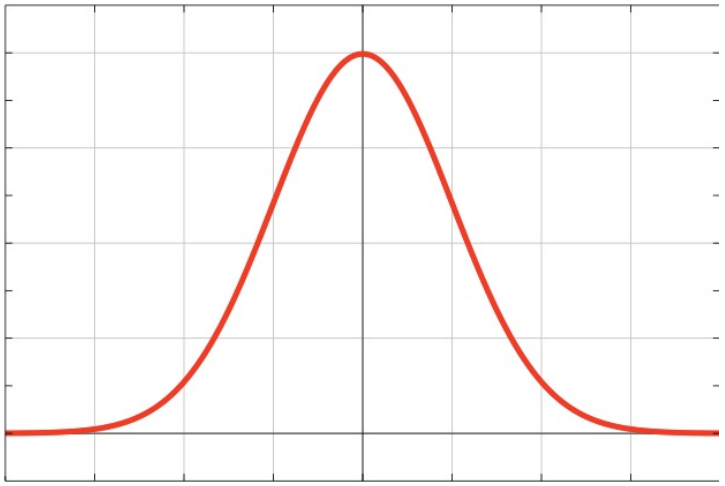
Traditional interpretation: View d as *cost* $\Rightarrow$

classic mass transportation problem *How transport one pile of sand (with a given mass density) into another prescribed density in a minimum cost way?* E.g., X uniform, $Y = f(X)$ Normal



f

$\longrightarrow$



Information theory/signal processing interpretation:

View d as *distortion*

How map/code/filter/transform one random object into another with a prescribed distribution in a way that makes the two as similar as possible *in the sense of minimizing the average distortion between them?*

Optimizing target distribution over some constraint set yields natural description of optimal code into a class of distributions:

$$d_{\text{Monge}}(\mu_X, \mathcal{P}) = \inf_{\nu \in \mathcal{P}} d_{\text{Monge}}(\mu_X, \nu)$$

Example: Quantization

M a positive integer

$\mathcal{P}_M = \{\text{all distributions on a finite subset } \hat{A} \subset A_X, |\ \hat{A}\ | \leq M\}$

$$d_{\text{Monge}}(\mu_X, \mathcal{P}_M) = \inf_{\hat{A}\subset A_X, |\hat{A}|\leq M} \inf_{\text{pmfs } \nu \text{ on } \hat{A}} \inf_{f: A_X \to \hat{A}, \nu = \mu_{f(X)}} E_{\mu_X}\left[\underbrace{d(X, f(X))}_{\geq \min_{y\in\hat{A}} d(X,y)}\right]$$

$$\geq \inf_{\hat{A}\subset A_X, |\hat{A}|\leq M} \underbrace{E_{\mu_X}\left[\min_{y\in\hat{A}} d(X, y)\right]}_{\equiv \Delta(\mu_X, \hat{A})} \equiv \Delta(\mu_X, M)$$

Mapping f of A_X into a finite subset is a *quantizer*

$$\Delta(\mu_X, M) = \inf_{\text{quantizers } q: |q(A_X)|\leq M} E_{\mu_X} d(X, q(X))$$

an operational distortion-rate function (DRF)

Converse Given finite $\hat{A}$, mapping $q^*_{\hat{A}}(x) = \text{argmin}_{y \in \hat{A}}\, d(x,y)$ is optimal & $\Rightarrow$ a distribution $\nu = \mu_{q^*(X)}$ on $\hat{A}$, hence

$$E_{\mu_X}\left[d(x, q^*_{\hat{A}}(x))\right] \geq \inf_{\text{pmfs } \nu \text{ on } \hat{A}} \quad \inf_{f: A_X \to \hat{A}, \nu = \mu_{f(X)}} E_{\mu_X}\left[d(X, f(X))\right]$$

Taking the infimum over all $\hat{A} \subset A_X :| \hat{A} | \leq M \Rightarrow$

$\Delta(\mu_X, M) \geq d_{\text{Monge}}(\mu_X, \mathcal{P}_M)$ $\Rightarrow$ $\boxed{d_{\text{Monge}}(\mu_X, \mathcal{P}_M) = \Delta(\mu_X, M)}$

Connects original Monge definition of distance with simple source coding problem of optimal quantization (Bennett, Lloyd, Shannon).

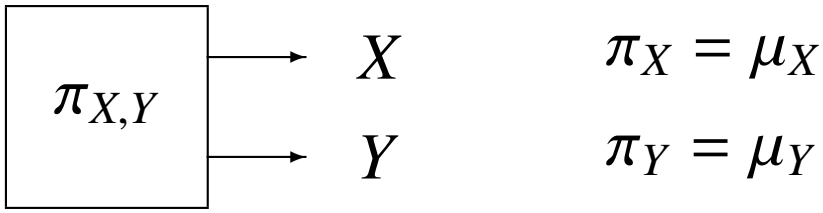
$\Delta(\mu_X, M)$ hard to compute, but $\exists$ good approximations for large M

Monge distance is hard to compute $\Rightarrow$ *Kantorovich generalization*

Kantorovich's problem: (1942): *Best* coupling

$$d_{\text{Kantorovich}}(\mu_X, \mu_Y) = \mathcal{T}(\mu_X, \mu_Y) \equiv \inf_{\pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\mu_Y} E_{\pi_{X,Y}}[d(X,Y)]$$

μ_X, μ_Y fixed. *Coupling* $\pi_{X,Y}$



Note: Set of such couplings is not empty,

e.g., product measure $\pi_{X,Y} = \mu_X \times \mu_Y$

Kantorovich generalized Monge's problem to a *linear programming* problem

Realized connection with Monge in 1948

Shared 1975 Nobel Economics Prize for invention of linear programming

Important question in literature: *When does Monge = Kantorovich?*

When is best stochastic coupling a deterministic mapping?

Milestones Monge (1781), **Kantorovich** (1942, metric distortion), Gini (1914), Salvemini (1943), Fréchet (1957), Dall Aglio (1956), Vasershtein/Wasserstein (1969, squared error), Dobrushin applied and popularized Vasershtein (1970), **Ornstein** (1970, average Hamming distance, *extension to processes*), Mallows (1972, squared error), Vallender (1973, metric^p, $p > 1$), Rubner and Guibas "earth mover's distance" (1998)

Outstanding surveys: Villani (2003, 2009), Rachev and Rüschendorf (1998). Villani has > 500 references

See also Rüschendorf : ☺

www.stochastik.uni-freiburg.de/~rueschendorf/papers/wasserstein.pdf

Back to quantization

Same argument works for Kantorovich d:

$$\begin{aligned}\mathcal{T}(\mu_X, \mathcal{P}_M) &= \inf_{\hat{A}\subset A_X, |\hat{A}|\le M} \inf_{\text{pmfs } \nu \text{ on } \hat{A}} \inf_{\pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\nu} E_{\pi_{X,Y}}\left[\underbrace{d(X,Y)}_{\ge \min_{y\in\hat{A}} d(X,y)}\right] \\ &\ge \inf_{\hat{A}\subset A_X, |\hat{A}|\le M} E_{\mu_X}\left[\min_{y\in\hat{A}} d(X,y)\right] = \Delta(\mu_X, M)\end{aligned}$$

$Y = q^*_{\hat{A}}(X) \Rightarrow \pi_{X,Y}$ with

$$\begin{aligned}E_{\pi_{X,Y}}[d(X,Y)] &= E_{\mu_X}\left[d(X, q^*_{\hat{A}}(X))\right] \\ &\ge \inf_{\text{pmfs } \nu \text{ on } \hat{A}} \inf_{\pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\nu} E_{\pi_{X,Y}}[d(X,Y)]\end{aligned}$$

Taking the infimum over $\hat{A}$ implies $\Delta(\mu_X, M) \ge \mathcal{T}(\mu_X, \mathcal{P}_M) \Rightarrow$

for Kantorovich as well as for Monge*

$$\Delta(\mu_X, M) = \mathcal{T}(\mu_X, \mathcal{P}_M) \qquad \star\, 1$$

Geometric representation for optimal quantization in terms of transportation distance

*Connection between best approximation in transportation sense and quantization developed in depth by Graf and Lushgy (2000)

Metric distortion

If underlying $d(x, y)$ on $A_X \times A_Y$ is a genuine distance (a metric), then much more is true thanks to the triangle inequality

If d is a metric, then $\mathcal{T}(\mu_X, \mu_Y)$ is also a metric and

$$\begin{aligned} \underbrace{\mathcal{T}(\mu_X, \mathcal{P}_M)}_{\Delta(\mu_X, M)} &= \inf_{\nu \in \mathcal{P}_M} \mathcal{T}(\mu_X, \nu) \leq \inf_{\nu \in \mathcal{P}_M} \left[\mathcal{T}(\mu_X, \mu_Y) + \mathcal{T}(\mu_Y, \nu)\right] \\ &= \mathcal{T}(\mu_X, \mu_Y) + \inf_{\nu \in \mathcal{P}_M} \mathcal{T}(\mu_Y, \nu) = \mathcal{T}(\mu_X, \mu_Y) + \mathcal{T}(\mu_Y, \mathcal{P}_M) \end{aligned}$$

$$\Rightarrow \boxed{|\, \Delta(\mu_X, M) - \Delta(\mu_Y, M) \,| \leq \mathcal{T}(\mu_X, \mu_Y) \qquad \star 2}$$

Optimal quantizer performance is *continuous* in the transportation distance.

More is true:

Quantizer mismatch and transportation

Given μ_X, μ_Y, coupling $\pi_{X,Y}$, finite $\hat{A}$,

$$\begin{aligned}\Delta(\mu_X, \hat{A}) &\equiv E_{\mu_X}\left(\min_{y\in\hat{A}} d(X,y)\right) \le E_{\pi_{X,Y}}\left(\min_{y\in\hat{A}}\left[d(X,Y)+d(Y,y)\right]\right) \\ &= E_{\pi_{X,Y}} d(X,Y) + E_{\mu_Y}\left(\min_{y\in\hat{A}} d(Y,y)\right)\end{aligned}$$

Taking the infimum over all couplings: If d is a metric

$$|\Delta(\mu_X, \hat{A}) - \Delta(\mu_Y, \hat{A})| \le \mathcal{T}(\mu_X, \mu_Y) \qquad \star 3$$

Useful bound in development of universal source codes and in proving consistency of empirical quantizer optimization (clustering) methods such as Steinhaus/Lloyd/k-means

[Steinhaus (1956), Lloyd (1957), MacQueen (1967)]

Distortion a power of a metric

Arguably most important distortion measure is squared error:
$d(x, y) = \|x - y\|^2$, *NOT* a metric

but it is the square of a metric!

Suppose $m(x, y)$ is underlying metric on $A_X \times A_Y$ and distortion $d(x, y) = m^p(x, y)$ for $p \geq 0$ transportation "distance" $\mathcal{T}_p$

$p = 0$ is Hamming distance, a metric

$0 \leq p \leq 1$, then $d(x, y)$ is still a metric $\Rightarrow$ previous results hold

Now focus on $p \geq 1$ most important case is $p = 2$

$$
\begin{aligned}
\Delta(\mu_X, \hat{A}) &= E_{\mu_X}\left(\min_{y\in\hat{A}} m^p(X,y)\right) = E_{\pi_{XY}}\left(\left[\min_{y\in\hat{A}} m(X,y)\right]^p\right) \\
&\leq E_{\pi_{XY}}\left(\left[\min_{y\in\hat{A}} (m(X,Y) + m(Y,y))\right]^p\right) \\
&= E_{\pi_{XY}}\left(\left[m(X,Y) + m(Y, q^*_{\hat{A}}(Y))\right]^p\right) \\
&\underbrace{\leq}_{\text{Minkowski}} \left\{\left[E_{\pi_{XY}}\left(m^p(X,Y)\right)\right]^{1/p} + \Delta(\mu_Y, \hat{A})^{1/p}\right\}^p
\end{aligned}
$$

Taking the infimum over all couplings: If $d = m^p, p \geq 1, m$ metric

$$
\boxed{|\,\Delta(\mu_X, \hat{A})^{1/p} - \Delta(\mu_Y, \hat{A})^{1/p}| \leq \mathcal{T}_p(\mu_X, \mu_Y)^{1/p}} \qquad \star\, 4
$$

extending mismatch and continuity to squared error

Summarize: If $d(x,y) = m^p(x,y)$ for m a metric and $p \geq 0$, then

$$\boxed{|\,\Delta(\mu_X, \hat{A})^{\min(1,1/p)} - \Delta(\mu_Y, \hat{A})^{\min(1,1/p)}| \leq \mathcal{T}_p(\mu_X, \mu_Y)^{\min(1,1/p)}}$$

$$\boxed{|\,\Delta(\mu_X, M)^{\min(1,1/p)} - \Delta(\mu_Y, M)^{\min(1,1/p)}| \leq \mathcal{T}_p(\mu_X, \mu_Y)^{\min(1,1/p)}}$$

Recall for general distortion measure (no metric assumptions)

$$\boxed{\Delta(\mu_X, M) = \inf_{\nu \in \mathcal{P}_M} \mathcal{T}(\mu_X, \nu)}$$

So far, *fixed-rate* quantization — transmit fixed $\log M$ bits for each input symbol giving index of codeword in codebook

Can get better distortion/rate tradeoff if use *variable-rate* quantization

Variable-rate quantization

More complicated. Need to decompose quantizer into component parts. Follow Chou et al. (1989), Linder (2002), focus on squared error.

Variable-length quantizer q consists of

- an *encoder* $\alpha : A_X \to \mathcal{I}$,
- a *decoder* $\beta : \mathcal{I} \to \hat{A}$ *(*reproduction codebook),
- and an *index coder* $\psi : \mathcal{I} \to \{0,1\}^*$, assumed to be a prefix-free code, with *admissible length function* $\ell(i)$ = length of codeword $\psi(i)$, $i \in \mathcal{I}$: $\sum_{i\in\mathcal{I}} 2^{-\ell(i)} \leq 1$ (Kraft inequality)

Output is $q(x) = \beta(\alpha(x))$ Instantaneous rate is $\ell(\psi(x))$

Average distortion: $E_{\mu_X}[d(X, \beta(\alpha(X)))] \equiv \Delta(\mu_X, q)$

Average rate: $E_{\mu_X}[\ell(\psi(X))] = L(\mu_X, q)$

Optimal performance: operational distortion-rate function

$$\Delta(\mu_X, R) = \inf_{q: L(\mu_X, q) \leq R} \Delta(\mu_X, q)$$

Common simplification: assume $\ell(i) = \ell^*(i) = -\log \Pr(\alpha(X) = i)$, in which case average rate is *Shannon entropy*

$$E_{\mu_X}[\ell(\alpha(X))] = -\sum_i \Pr(\alpha(X) = i) \log \Pr(\alpha(X) = i) \equiv H(\alpha(X)) = H(q(X))$$

For a random variable Y with a discrete distribution ν,

$$H(Y) = H(\nu) \equiv -\sum_y \nu(y) \log \nu(y)$$

Entropy-constrained quantization

Optimal entropy-constrained quantizer performance: operational DRF

$$\Delta(\mu_X, R) = \inf_{q: H(q(X)) \le R} \Delta(\mu_X, q)$$

Reasons for focus on entropy-constrained:

1. If allow non integer lengths, ℓ^* is the *optimal length function* over all admissible length functions: $E_{\mu_X}[\ell(\alpha(X))] \ge H(\alpha(X))$
2. Approximate equality under suitable conditions
3. Easier to compute than finding optimal index coder

Under suitable assumptions

$$\Delta(\mu_X, R) = \inf_{\nu: H(\nu) \le R} \mathcal{T}(\mu_X, \nu) \qquad \star 5$$

Lagrangian formulation

But this approach does not yield deeper connections of quantization with transportation

Alternative approach is to replace constrained optimization by Lagrangian optimization

Resonates with other threads of transportation/quantization connection

Constrained optimization for fixed R *almost* the same as unconstrained Lagrangian optimization

$$\Delta(\mu_X, \lambda) = \inf_q \left(\Delta(\mu_X, q) + \lambda H(q(X))\right)$$

for some $\lambda > 0$

(only get convex hull of R solutions, but usually enough)

$$\lambda \to \infty \Leftrightarrow R \to 0, \lambda \to 0 \Leftrightarrow R \to \infty$$

Analagous to fixed-rate results, Linder (2002) showed that for squared error

$$\Delta(\mu_X, \lambda) = \inf_{\nu} \left[\mathcal{T}_2(\mu_X, \nu) + \lambda H(\nu)\right] \qquad \star 6$$

where the infimum is over all discrete distributions ν with finite second moment and finite entropy. Since $H(\nu)$ depends only on ν,

$$\begin{aligned}
\Delta(\mu_X, \lambda) &= \inf_{\nu} \left[\mathcal{T}_2(\mu_X, \nu) + \lambda H(\nu)\right] \\
&= \inf_{\nu} \left(\inf_{\pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\nu} E_{\pi_{X,Y}} d(X,Y) + \lambda H(\nu) \right) \\
&= \inf_{\nu} \left(\underbrace{\inf_{\pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\nu} E_{\pi_{X,Y}} \left[d(X,Y) - \lambda \log \nu(Y)\right]}_{\equiv \mathcal{T}(\mu_X,\nu;\lambda)} \right)
\end{aligned}$$

$\mathcal{T}(\mu_X, \nu; \lambda)$ can be viewed as a *penalized* or *regularized* variation on the transportation distance: $H(\nu)$ can be considered as a measure of the cost or complexity of the second argument. Provides extension of $(\star 1)$ to variable-rate quantization:

$$\Delta(\mu_X, \lambda) = \inf_{\nu} \mathcal{T}(\mu_X, \nu; \lambda) \qquad \star 7$$

Similar ideas of regularized transportation have been explored in different contexts in the transportation literature.

Linder (2002) showed that the optimal performance mismatch result remains true for the variable-rate case and squared error, with the ordinary L_2 transportation distance:

$$|\, \Delta(\mu_X, \lambda)^{1/2} - \Delta(\mu_Y, \lambda)^{1/2}| \leq \mathcal{T}_2(\mu_X, \mu_Y)^{1/2} \qquad \star 8$$

$\mathcal{T}(\mu_X, \mu_Y; \lambda)$ suggests another transportation distance if replace Shannon entropy by Shannon mutual information:

Shannon Mutual Information

Given jointly distributed discrete random variables X, Y with distribution $\mu_{X,Y}$:

$$I(X;Y) \equiv I(\mu_{X,Y}) = H(X)+H(Y)-H(X,Y) = \sum_{x,y} \mu_{X,Y}(x,y) \log \frac{\mu_{X,Y}(x,y)}{\mu_X(x)\mu_Y(y)}$$

In general case, $I(X,Y) \equiv \sup_{\text{all quantizers } q,r} I(q(X), r(Y))$

Constructive description: if $\mu_{X,Y} \ll \mu_X \times \mu_Y$, then *information density*

$$i \equiv \log \frac{d\mu_{X,Y}}{d\mu_X \times \mu_Y} \text{ exists and } I(X;Y) = E_{\mu_{X,Y}} i(X,Y)$$

Key fact: $I(X;Y) \leq H(X) = I(X;X) \Rightarrow$

$$\begin{aligned}\Delta(\mu_X, R) &= \inf_{q:H(q(X))\leq R} E_{\mu_X} d(X, q(X)) \geq \inf_{\pi_{X,Y}:\pi_X=\mu_X, H(Y)\leq R} E_{\pi_{X,Y}} d(X,Y) \\ &\geq \inf_{\pi_{X,Y}:\pi_X=\mu_X, I(X,Y)\leq R} E_{\pi_{X,Y}} d(X,Y) \equiv D(\mu_X, R),\end{aligned}$$

Shannon distortion-rate function

Most evaluations use Lagrangian approaches: Instead of $D(\mu_X, R)$ find

$$D(\mu_X, \lambda) = \inf_{\pi_{X,Y}:\pi_X=\mu_X} E_{\pi_{X,Y}} \left[d(X,Y) + \lambda i(X,Y)\right]$$

Analogy with transportation & previous results suggests

$$D(\mu_X, \lambda) \quad = \quad \inf_{\nu} \left\{ \underbrace{\inf_{\pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\nu} E_{\pi_{X,Y}} \left[d(X,Y) + \lambda i(X,Y) \right]}_{\mathcal{T}^{(\mathsf{i})}(\mu_X,\nu;\lambda)} \right\}$$

Term in curly brackets suggests a transportation distance with alphabet-level distortion $d_\lambda(x, y) = d(x, y) + \lambda i(x, y)$.

Not a valid distortion since can be negative, but its expectation is ≥ 0!

of any use apart from representation for Shannon DRF?

So far: all "one-shot" source coding — scalars or vectors

Process Transportation Distance

Huge literature on theory and applications of optimal transportation to many branches of mathematics, mostly for random variables/vectors — *not processes*.

There is a literature on coupling random processes, but

not in context of optimal transportation.

Optimal transportation extended to processes in 1970s in ergodic theory and information theory by Donald S. Ornstein and his colleagues ($\overline{d}$ or d-bar distance)

Ornstein won the 1974 Bôcher Memorial Prize for his proof of the isomorphism theorem for *Bernoulli shifts*

Need notion of distortion between variables and vectors of all sizes. Shannon *fidelity criterion* does this:

$$d_N(x^N, y^N) \text{ on } A_X^N \times A_Y^N,\ N = 1, 2, \ldots$$

Usually assume additive: $d_N(x^N, y^N) = \sum_{k=0}^{N-1} d(x_k, y_k)$

Define process transportation "distance" by

$$\overline{\mathcal{T}}(\mu_X, \mu_Y) \equiv \sup_N N^{-1} \mathcal{T}(\mu_{X^N}, \mu_{Y^N})$$

$\overline{\mathcal{T}}_0$ is Ornstein's (1970) $\overline{d}$ (average Hamming), $\overline{\mathcal{T}}_2$ is the ρ-bar distance (squared error) (1975)

Basic Properties

μ_X, μ_Y stationary with common alphabet, additive fidelity criterion, per-symbol d = metric^p, $p \geq 0$

- $\overline{\mathcal{T}}_p(\mu_X, \mu_Y) = \lim_{N\to\infty} N^{-1}\mathcal{T}_p(\mu_{X^N}, \mu_{Y^N})$ supremum = limit
- $\overline{\mathcal{T}}_p(\mu_X, \mu_Y) = \inf_{\text{stationary } \pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\mu_Y} E_{\pi_{X,Y}}[d(X_0, Y_0)]$ direct process definition
- $\overline{\mathcal{T}}_p^{\min(1,1/p)}(\mu_X, \mu_Y)$ is a *metric* on stationary processes
- If μ_X, μ_Y are both stationary and ergodic, then

  $\overline{\mathcal{T}}_p(\mu_X, \mu_Y)$ = *amount by which a μ_X-frequency-typical sequence must be changed in the time-average d sense in order to confuse it with a μ_Y-frequency-typical sequence*.

- If μ_X, μ_Y are both stationary and ergodic,

$$\overline{\mathcal{T}}_p(\mu_X, \mu_Y) = \inf_{\text{stationary \& ergodic } \pi_{X,Y}:\pi_X=\mu_X,\pi_Y=\mu_Y} E_{\pi_{X,Y}}\left[d(X_0.Y_0)\right]$$

- If μ_X, μ_Y are both IID, then $\overline{\mathcal{T}}_p(\mu_X, \mu_Y) = \mathcal{T}_p(\mu_{X_0}, \mu_{Y_0})$

  $\Rightarrow$ If both processes IID, simple random variable transportation enough

---

Examples:

- IID coin flips with bias p, q: $\mathcal{T}_0(\mu_X, \mu_Y) = \mid p - q \mid$
- Zero mean Gaussian with power spectral densities $S_X(f)$ and $S_Y(f)$: $\mathcal{T}_2(\mu_X, \mu_Y) = \int \mid \sqrt{S_X(f)} - \sqrt{S_Y(f)} \mid^2 df$

in both cases Monge = Kantorovich

Process Information Rates

X stationary, process distribution μ_X

Entropy rate

Discrete alphabet $\overline{H}(X) = \overline{H}(\mu_X) = \lim_{N\to\infty} \frac{H(X^N)}{N}$

General alphabet $\overline{H}(X) = \sup_{\text{quantizers } q} \overline{H}(q(X))$

Mutual Information rate:

Discrete alphabet $\overline{I}(X;Y) = \overline{I}(\mu_{X,Y}) = \lim_{n\to\infty} \frac{1}{N} I(X^N;Y^N)$

General alphabet $\overline{I}(X;Y) = \sup_{\text{quantizers } q,r} \overline{I}(q(X);r(Y)).$

Shannon distortion-rate function for a stationary process is

$$D(\mu_X, R) = \lim_{N\to\infty} \frac{1}{N} D(\mu_{X^N}, R) = \inf_{\pi_{X,Y}:\pi_X=\mu_X, \bar{I}(X,Y)\le R} E_{\pi_{X,Y}} d(X_0, Y_0)$$

Transportation and Entropy

If processes have common finite alphabet A with $|\ A\ |$ symbols,
Fano's inequality for processes $\Rightarrow$

$$|\ \overline{H}(\mu_X) - \overline{H}(\mu_Y)\ | \leq \overline{d}(\mu_X, \mu_Y) \log(|\ A\ | - 1) + h_2(\overline{d}(\mu_X, \mu_Y))$$

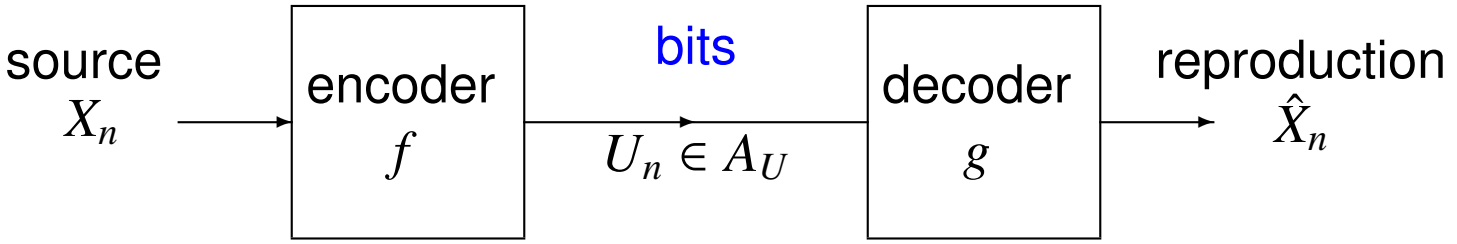
where $h_2(x) = -x \log x - (1 - x) \log(1 - x)$

Entropy rate continuous with respect to $\overline{d}$

Marton's inequality relating $\overline{d}$ distance to relative entropy rate $\Rightarrow$

if one process IID equiprobable, then converse is true.

Source Coding and Transportation



Consider stationary (sliding-block) codes f, g

Average distortion $\Delta(\mu_X, f, g) = E[d(X_0, \hat{X}_0)]$

Operational DRF for process $\Delta(\mu_X, R) \equiv \inf_{f,g:\log|A_U|\leq R} D(\mu_X, f, g)$

Shannon source coding theory $\Rightarrow$

$$\Delta(\mu_X, R) = D(\mu_X, R) = \inf_{\nu:\overline{H}(\nu)\leq R} \overline{\mathcal{T}}_p(\mu_X, \nu)$$

Implications

For any metric $d = m^p$ with m metric and $p \geq 0$,

- Optimal performance using quantizers, vector quantizers (block source codes), and stationary codes is continuous in $\overline{\mathcal{T}}_p$
- Shannon distortion-rate functions are continuous in $\overline{\mathcal{T}}_p$
- Shannon rate-distortion functions are continuous in $\overline{\mathcal{T}}_p$

  (except possibly at $R = 0$)
- Can extend transportation distance incorporating entropy or mutual information to processes

Final thoughts

- Brief survey of optimal transportation from a quantization viewpoint
- Connections between and combinations of transportation distance and Shannon information
- Geometric view of source coding as minimum transportation distance selection of entropy constrained model and as minimum unconstrained Lagrangian distance selection
- Lagrangian transportation combining distance/distortion + information
- Suppose $|\mathcal{P}| \leq M$. Then $\operatorname{argmin}_{\nu \in \mathcal{P}} \mathcal{T}(\mu_X, \nu)$ = a minimum transportation distance classifier, can cluster with respect to $\mathcal{T}$