Information, concentration, transportation of exponentially concave functions

Soumik Pal
University of Washington, Seattle
(Mostly) Joint work with T.-K. Leonard Wong

May 11, 2016

Part I: Transportation

Definition

- $\Delta_n \subseteq \mathbb{R}^n$. Convex.
- $\varphi : \Delta_n \to \mathbb{R} \cup \{-\infty\}$ is exponentially concave if e^{φ} is concave.

$$\mathrm{Hess}(\varphi) + \nabla\varphi\,(\nabla\varphi)' \leq 0.$$

- Throughout Δ_n is the open unit simplex.

$$\Delta_n := \left\{ (p_1, \ldots, p_n) : \; p_i > 0, \; \sum_i p_i = 1 \right\}.$$

- Examples: $p, \pi \in \Delta_n$, $0 < \lambda < 1$.

$$\begin{aligned} \varphi(p) &= \frac{1}{n}\sum_i \log p_i, \quad \varphi(p) = \sum_i \pi_i \log p_i, \\ \varphi(p) &= \log\left(\sum_i \pi_i p_i\right), \quad \varphi(p) = \frac{1}{\lambda}\log\left(\sum_i p_i^{\lambda}\right). \end{aligned}$$

Several recent occurrences

- Stochastic portfolio theory (will discuss). Fernholz '99, '02.
- Entropic Curvature-Dimension conditions and Bochner's inequality. Erbar, Kuwada, and Sturm '15.
- Proof of the Fundamental Gap Conjecture. Andrews and Clutterbuck '11.
- Statistics, optimization, machine learning. Cesa-Bianchi and Lugosi '06, Mahdavi, Zhang, and Jin '15.
- Unified study is lacking. Compare log-concave functions.

Gradients of e-concave functions

- Fact 1: Gradients of e-concave functions are probabilities.
- (Fernholz '02, P. and Wong '15). φ, e-concave on Δ_n. Define π by
$$\pi_i = p_i \left(1 + D_{e(i)-p}\varphi(p)\right).$$
Then $\pi \in \Delta_n$. $e(i)$ is ith standard basis vector.
- Portfolio map: $\pi : \Delta_n \to \Delta_n$.
- Example: $\varphi(p) = \frac{1}{n}\sum_i \log p_i$. Then $\pi(p) \equiv (1/n, \ldots, 1/n)$.

Gradients of e-concave functions

- Fact 3: Gradients of e-concave functions are optimal transports.
- Recall Monge-Kantorovich OT problem.
- $\mathcal{X} = \overline{\Delta_n}$, $\mathcal{Y} = [-\infty, \infty)^n$.

$$c(p, \gamma) = \log \sum_{i=1}^{n} e^{\gamma_i} p_i = \log \int e^{\gamma} dp.$$

- Consider probabilities P on $\mathcal{X}$ and Q on $\mathcal{Y}$.

$$\inf \mathrm{E}_R \left(c(p, \gamma) \right), \quad \text{over all couplings } R \text{ of } (P, Q).$$

- Solution is an optimal coupling $(p, \gamma(p))$.
- Also see Oliker '07 for the Alexandrov problem on the sphere.

Exponential change of measures

Theorem (P.-Wong '15)

Consider optimal coupling $(p, \gamma(p))$ *for some* (P, Q). *Let*

$$\pi_i = \frac{e^{\gamma_i} p_i}{\sum_j e^{\gamma_j} p_j}, \quad i = 1, \ldots, n.$$

Then $\pi = \pi(p)$ *is the gradient of some e-concave function* φ *on* Δ_n.
And conversely ...

Compare convex functions whose gradients are optimal transports for cost $\|x - y\|^2$.

Exponential coordinate systems

- Useful to see Δ_n as a manifold with exponential coordinates $(\theta_1, \ldots, \theta_{n-1}) \in \mathbb{R}^{n-1}$.

$$\theta_i = \log\left(p_i/p_n\right), \quad p_i = \exp\left(\theta_i - \psi(\theta)\right), \quad i = 1, \ldots, n-1.$$

  Here

$$\psi(\theta) = \log\left[1 + \sum_i e^{\theta_i}\right].$$

- Consider P, Q on $\mathbb{R}^{n-1}$ and $c(\theta, \phi) = \psi(\theta - \phi)$.
- MK OT problem is exactly as before.
- Exponential coordinate of $\pi(p(\theta)) = \theta - \phi(\theta)$.

Example

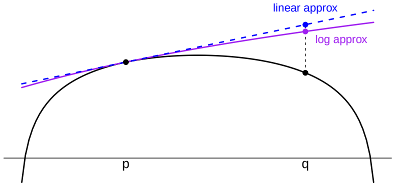
- 3 coordinates: $p \leftrightarrow \theta \leftrightarrow \phi$.
- θ - primal coordinate. ϕ - dual coordinate.
- Example: $P(\theta) = \otimes N(a_i, \sigma_i^2)$. $Q(\phi) = \otimes N(b_i, (1-\lambda)\sigma_i^2)$.
- Optimal coupling:

$$\varphi(p) = \frac{1}{\lambda} \log \left(\sum_j c_j p_j^\lambda \right), \quad \pi_i(p) = \frac{c_i p_i^\lambda}{\sum_j c_j p_j^\lambda}.$$

Here $c \in \Delta_n$ are given by

$$(1-\lambda)a_i - \log \frac{c_i}{c_n} = b_i, \quad \text{for al } i.$$

L-divergence vs Bregman divergence



- Fact 3: (P.- Wong '15) e-concave functions have better approximation by gradient. L-divergence:

$$T(q \mid p) = \log(1 + \nabla\varphi(p) \cdot (q - p)) - (\varphi(q) - \varphi(p)) \geq 0.$$

- (Classical information geometry) Bregman divergence

$$D(q \mid p) = \nabla\varphi(p) \cdot (q - p) - (\varphi(q) - \varphi(p))$$

L-divergence vs Bregman divergence

- $\varphi(p) = -\sum_i p_i \log p_i$. Shannon entropy. Convex.

$$D(q \mid p) = \sum_i q_i \log \frac{q_i}{p_i}, \quad \text{Relative entropy.}$$

- Fix $\pi \in \Delta_n$. $\varphi(p) = \sum_i \pi_i \log p_i$. Cross entropy. e-concave.

$$T(q \mid p) = \log\left(\sum_i \pi_i \frac{q_i}{p_i}\right) - \sum_i \pi_i \log\left(\frac{q_i}{p_i}\right).$$

- Free energy.

Part II: Information

A wee bit of mathematical finance

- Optimal transports are closely connected to economics.
- Market weights for n stocks:
- μ_i = Proportion of the total capital that belongs to ith stock.
- Process in time, $\mu(t),\ t = 0, 1, 2, \ldots$ in Δ_n.
- Portfolio: $\pi = (\pi_1, \ldots, \pi_n) \in \Delta_n$. Process in time $\pi(t)$.
- Portfolio weights:

$$\pi_i = \text{Proportion of the total value that belongs to } i\text{th stock.}$$

- For us $\pi = \pi(\mu) : \Delta_n \to \overline{\Delta_n}$.

Relative value

- How does the portfolio π compare with an index, say, S&P 500?
- Start by investing \$1 in portfolio and compare with index.
- Relative value process: $V(\cdot) =$ ratio of growth of \$1.

$$\frac{\Delta V(t)}{V(t)} = \sum_{i=1}^{n} \pi_i(t) \frac{\Delta \mu_i(t)}{\mu_i(t)}, \quad V(0) = 1.$$

  Discrete exponential integral.
- How does $V(\cdot)$ behave?

Fernholz decomposition

- Suppose π is generated by an e-concave function.
- (Fernholz '99, P. and Wong '15)

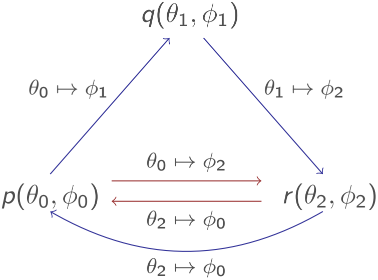
$$\log V(t) = \varphi(\mu(t)) - \varphi(\mu(0)) + \sum_{s=0}^{t-1} T\left(\mu(s+1) \mid \mu(s)\right).$$

- Here, L-divergence:

$$\begin{aligned} T(q \mid p) &= \log\left(\sum_i \pi_i(p)\frac{q_i}{p_i}\right) - \sum_i \pi_i(p)\log\frac{q_i}{p_i} \\ &= \log(1 + \nabla\varphi(p)\cdot(q-p)) - (\varphi(q) - \varphi(p)) \geq 0. \end{aligned}$$

- $\lim_t \log V(t) = \infty$ if $\varphi(\mu(t))$ is bounded. Volatility harvesting.
- (P. and Wong '15). π has above property $\Rightarrow$ π is solution of OT.

Main question

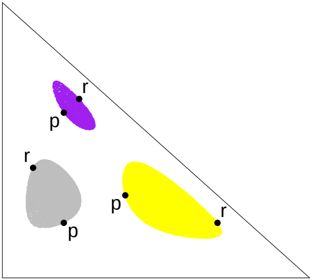


- Given φ e-concave, can I characterize $(p, q, r) \in \Delta_n^3$ such that

$$T(q \mid p) + T(r \mid q) \leq T(r \mid p)?$$

- Determines optimal frequency of trading the portfolio.
- Has the optimal transport interpretation as above.

Plots of $\{q \in \Delta_n : T(q \mid p) + T(r \mid q) \leq T(r \mid p)\}$



- (P. and Wong '16) Take any q on boundary. Then (p, q, r) forms a "right angle triangle".
- The sides are geodesics of a geometry and angles are given by a Riemannian metric.

Information geometry (Amari and Nagaoka (1982))

- Such questions have been studied for relative entropy.
- Exponential family Δ_n

$$p(x, \theta) = e^{\sum_i \theta_i \delta_i(x) - \psi(\theta)}, \quad x \in \{1, \dots, n\}.$$

- Metric is Fisher information:

$$g_{ij}(\theta) = -\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(X, \theta) \right]$$

  Riemannian metric on Δ_n
- Flat geometry for any Bregman divergence

A new information geometry

- Riemannian metric g: u, v in tangent bundle at p.

$$\langle u, v \rangle_p \quad = \quad -v^T \left(\text{Hess } \varphi(p) + \nabla\varphi(p)\nabla\varphi(p)^T \right) u.$$

- Define dual connection ∇^* using dual variables $\theta \mapsto \phi$.
- Best described by geodesics: $\gamma^* : [0, 1] \to \Delta_n$ curves

$$\nabla^*_{\dot{\gamma}^*(t)} \dot{\gamma}^*(t) = 0 \quad \text{dual geodesic.}$$

- Parallel transport of tangents.

Geodesics and curvatures

Theorem (Pal and W. (2016))

For a suitable ∇^* the following are true. For $p \in \Delta_n$, consider dual coordinate ϕ from OT problem. Define p^* by

$$p(\theta, \phi) \mapsto p^* \in \Delta_n, \quad p_i^* = e^{-\phi_i - \psi(-\phi)}$$

(i) *If $\gamma^*(t)$ is a dual geodesic, its image $p^*(t)$ is a Euclidean straight line in Δ_n.*

(ii) *Dual geodesics are gradient flows.*

$$\begin{aligned} \dot{\gamma}^*(t) &= -\mathrm{grad}\, T(\cdot \mid p)(\gamma^*(t)), \quad \gamma^*(0) = q \\ \left(\exp_q^*\right)^{-1}(p) &\propto -\mathrm{grad}\, T(\cdot \mid p)(q). \end{aligned}$$

In particular, the induced geometry is not flat. Moreover, Δ_n has constant dual (sectional) curvature -1. In particular, $\mathrm{Ric}^* = -(n-2)g$.

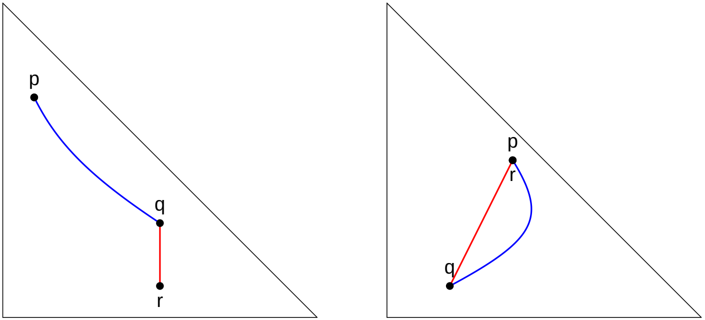
Generalized Pythagorean theorem

Theorem (P. and Wong (2016))

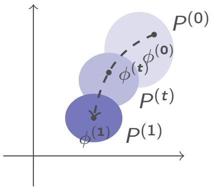
For any $p, q, r \in \Delta_n$, the equality

$$T(q \mid p) + T(r \mid q) = T(r \mid p)$$

holds if and only if the dual geodesic joining q and p and the Euclidean straight line joining q and r meet orthogonally at q.



Geodesics and displacement interpolation



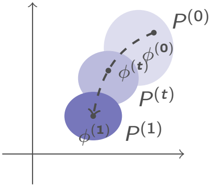
- Let $P^{(0)}, P^{(1)}$ be Borel probability measures on $\mathbb{R}^{n-1}$.
- Transport problem be solved in terms of the exponentially concave function φ.
- Define a cost on the paths of transporting $\theta \mapsto \phi$ by a Lagrangian action.

$$\psi(\theta - \phi) = \inf \left\{ \int_0^1 - \log \left(\frac{1}{n} + \frac{d}{dt} e^{-\psi(\gamma(0) - \gamma(t))} \right) dt : \gamma(0) = \theta, \gamma(1) = \phi \right\}.$$

$$\psi(\theta) = \log \left[1 + \sum_i e^{\theta_i} \right].$$

- Cost minimizing paths? For convex functions, leads to a beautiful theory of Otto, Lott, Villani in Wasserstein space where cost is $\|\cdot\|^2$.

Displacement interpolation



Theorem (P. and Wong '16)

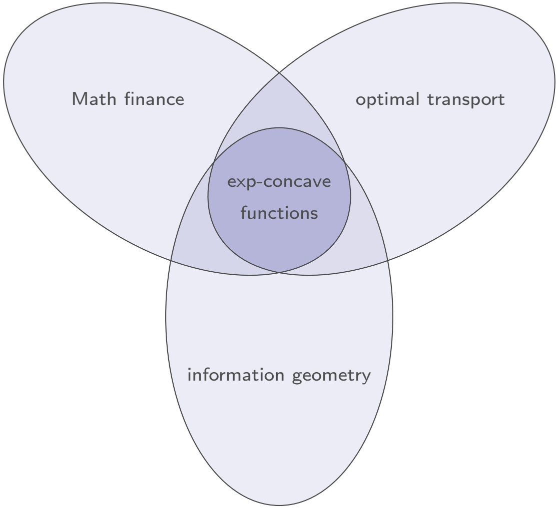
(i) *The cost minimizing paths are dual geodesics joining* (θ, ϕ).

(ii) *(Intermediate time optimality) For each* $t \in [0, 1]$, *the coupling* $(\theta, \phi^{(t)}(\theta))$ *solves the transport problem for* $(P^{(0)}, P^{(t)})$ *where* $P^{(t)}$ *is the push-forward.*

(iii) *For* $t \in [0, 1]$, *the portfolio map is* $\pi^{(t)} = (1 - t)\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) + t\pi$.

(iv) *Generated by e-concave function*

$$\varphi_t(p) = \frac{(1-t)}{n} \sum_i \log p_i + t\varphi(p).$$

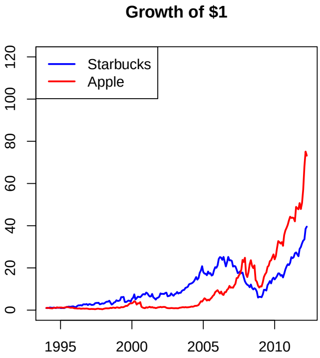

Math finance
optimal transport
exp-concave
functions
information geometry

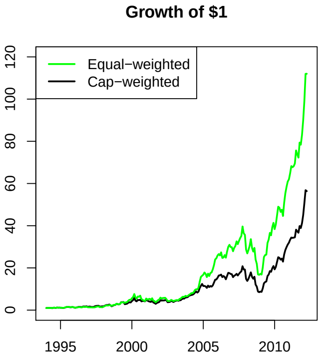
Concentration

Welcome to the dismal science of economics

- Most institutional investors (e.g. Vanguard) use ETF.
- Tracks a market index such as S&P 500.
- CAPM and Efficient Market Hypothesis support the idea.
- Nobel prize 2013 - Fama and Shiller.
- Portfolios from e-concave functions do better than the index without statistical assumptions.
- See P. and Wong '15. A model free notion of volatility.

How do these portfolios work in practice?





One can do much better with reduced risk by using concentration of measure.

Exp concavity in high dimensions

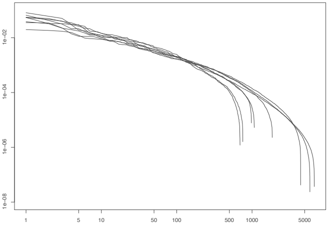
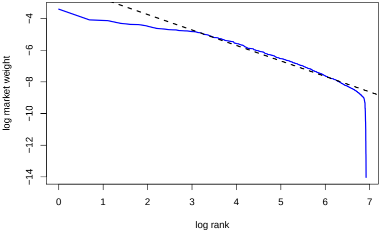


Figure 1: Capital distribution curves: 1929–1999



- Ordered market weights are typically Pareto (polynomially decaying):

$$\mu_{(1)}(t) \geq \mu_{(2)}(t) \geq \ldots \geq \mu_{(n)}(t), \quad \log \mu_{(i)} \propto i^{-\alpha}.$$

- Slope $\alpha \in [-1.2, .08]$. Axtell '01 *Science*.

The Pareto distribution

- Fix $\alpha \in [1/2, 1]$. Define $\nu^{(n)} \in \Delta_n$ by

$$\nu_i^{(n)} = \frac{i^{-\alpha}}{\sum_{j=1}^{n} j^{-\alpha}}.$$

- Can generalize to regularly varying sequences with index in $[1/2, 1]$.
- Consider Dirichlet distribution $\mathrm{Dir}\left(n\nu^{(n)}\right)$.
- Assumption 1: $\left\|\mu(0) - \nu^{(n)}\right\|$ has the same distribution as $\mu(0) \sim \mathrm{Dir}(n\nu^{(n)})$.
- Assumption 2: μ is a continuous semimartingale process that is "slow to escape $O(1/\sqrt{n})$ neighborhoods of $\nu^{(n)}$".

(K,N) exponential concavity (Erbar-Kuwada-Sturm '14)

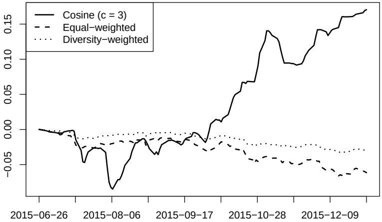
- Define $(n, 1)$ e-concave function on Δ_n.

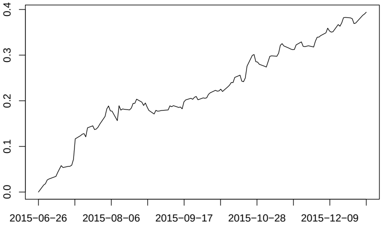
$$\varphi(p) = \log\cos\left(\sqrt{n}\left\|p - \nu^{(n)}\right\|\right), \quad \left\|p - \nu^{(n)}\right\| < \frac{\pi}{2\sqrt{n}}.$$

- Unit simplex under Dir $\left(n\mu^{(n)}\right)$ has diameter $1/\sqrt{n}$.
- (P. '16) $\exists\, g_n = O(n^{\alpha - 1/2})$ such that

$$P\left(\log V(1/\sqrt{\log n}) \geq g_n\right) = 1 - O\left(\exp\left(-c_0 n^{(1-\alpha)/4}\right)\right).$$

Cosine portfolios





- $n = 1000$. $\alpha \in [0.75, 0.95]$. Jun - Dec 2015.
- Distance from Pareto scales like $\sqrt{n}$.
- Cosine portfolios generated by $(n, 1)$ e-concave function.
- Beats the index by 15% in 6 months.

Many thanks to the organizers and all the participants