Statistical inference of convex order by Wasserstein projection

Jakwang Kim
PIMS Kantotrovich Initiative, University of British Columbia
Joint work with: Young-Heon Kim(UBC), Yuanlong Ruan(Beihang University), Andrew Warren(UBC)

SAARC Optimal Transport and applications
June 2, 2024

Outline

Convex order

Stochastic order

Suppose you are a gambler, and there are two gambles, say, blackjack and roulette. To get more money (or to lose less money in fact), which gamble should you choose? Which one is better than the other?
Blackwell[1] suggested a criterion to compare two different random variables: called stochastic dominance.
Let $X \sim \mu$ and $Y \sim \nu$ with finite first moments. We say that Random variable X has first-order stochastic dominance over random variable Y if for any $t \in \mathbb{R}$

$$\mathbb{P}(X \geq t) \geq \mathbb{P}(Y \geq t)$$

and a strict inequality holds for some t. Equivalently, for any non decreasing $u : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}u(X) \geq \mathbb{E}u(Y).$$

[1] Blackwell, "Equivalent comparisons of experiments".

Stochastic order

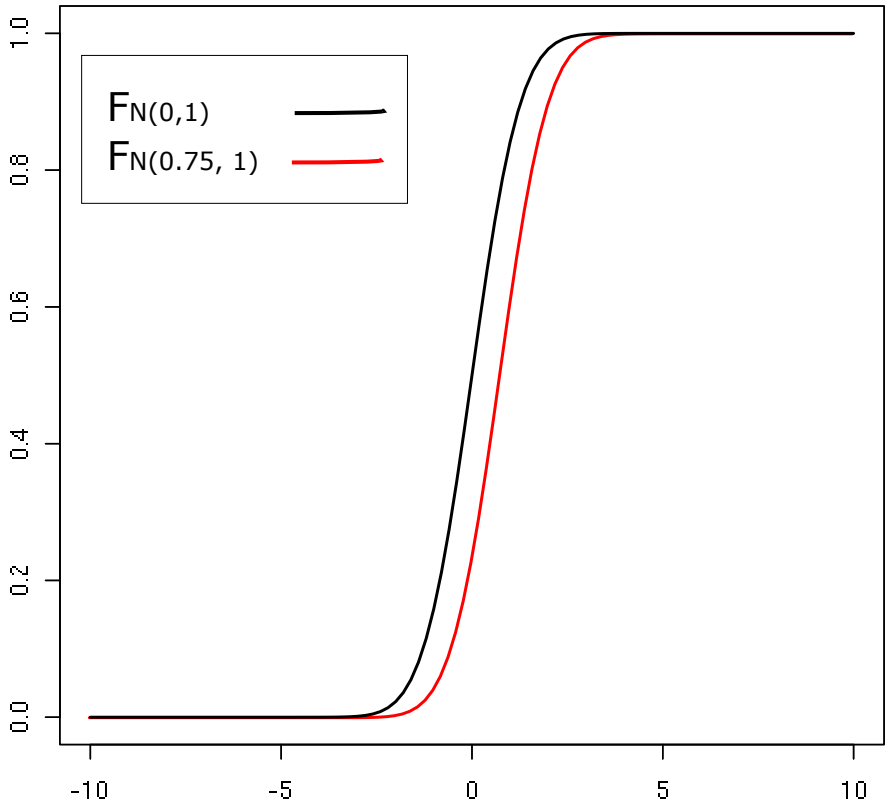


Figure: $X \sim N(0.75, 1)$ and $Y \sim N(0, 1)$: X dominates Y.

Stochastic order

People in other areas, economics, finance, actuarial science, operation research etc., in which decision making under uncertainty is of interest has realized its importance and usefulness. (Davidson and Duclos[2], Rothschild and Stiglitz[3], Linton, Post and Whang[4])
More generally, one can define (a variant) stochastic order by choosing different *defining class*. Let $\mathcal{A}$ be a defining class, i.e. a certain family of functions: e.g. $\mathcal{A} = \{\text{non-decreasing}\}$ for first-order stochastic dominance. We say that X dominates Y with $\mathcal{A}$, and denoted by $Y \preceq_{\mathcal{A}} X$ if for every $\varphi \in \mathcal{A}$,

$$\mathbb{E}\varphi(Y) \leq \mathbb{E}\varphi(X).$$

See Belzunce, Martínez-Riquelme and Mulero[5] for more details.

---

[2] Davidson and Duclos, "Statistical inference for stochastic dominance and for the measurement of poverty and inequality".

[3] Rothschild and Stiglitz, "Increasing risk: I. A definition".

[4] Linton, Post, and Whang, "Testing for the stochastic dominance efficiency of a given portfolio".

[5] Belzunce, Martínez-Riquelme, and Mulero, *An introduction to stochastic orders*.

Convex order

Convex order is a stochastic order with a defining class consisting of convex functions. Formally, we say that $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are in convex order, or μ is dominated by ν in convex order, denoted by $\mu \preceq \nu$, if for any convex function φ,

$$\mathbb{E}_\mu \varphi(X) \leq \mathbb{E}_\nu \varphi(Y).$$

Notice that if μ and ν are in convex order, by constant functions($+1$ and -1), they have the same mean.

Convex order

It is known by Strassen[6] that μ and ν are in convex order if and only if there is a martingale coupling between them, i.e., there is a joint distribution $\pi \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ such that whose marginals are μ and ν, respectively and

$$\mathbb{E}_{(X,Y)\sim\pi}[Y|X] = X$$

μ-almost everywhere.

[6] Strassen, "The existence of probability measures with given marginals".

Some Motivations

Model-free option pricing in finance

Let us consider the problem of pricing forward start options at time 0. We want to price a security payment $p(X, Y)$ where X and Y are the forward price vectors at time 1 and 2, respectively. Rather than assuming any specific model, financial mathematicians want to know the call prices at 1 and 2 with no-arbitrage. No-arbitrage is equivalent to $\mathbb{E}[Y|X] = X$, martingale condition. Since knowing European call option prices (for the continuum of strikes) is equivalent knowing the marginal distributions under a risk-neutral measure[7], without loss of generality we assume that $X \sim \mu$ and $Y \sim \nu$ for some known μ and ν. The goal is to find a martingale (no-arbitrage) coupling between μ and ν which solves

$$\sup_{\mathbb{E}[Y|X]=X} \mathbb{E}\left[p(X, Y)\right].$$

[7] Breeden and Litzenberger, "Prices of state-contingent claims implicit in option prices".

Model-free option pricing in finance

A risk-neutral martingale coupling is not observable but instead marginally obtained via calibrations to option prices trading on the market. In reality, the procedure can be hampered by practical reasons. Thus a good extracting procedure should address these difficulties and importantly be consistent across maturities, i.e. $\mu \preceq \nu$. Clearly being able to statistically test the convex order relationship between μ and ν would be useful for the task of tuning an effective extracting procedure. Additionally, large deviation from the expected convex order that $\mu \preceq \nu$ would imply existence of arbitrage.

Labor economics

The central question in labor economics is how workers are matched to firms in such a way that productivity is maximized. A worker may possess different or multidimensional skills, while a firm may be interested in only a fraction of the various skills of a worker, but it can only hire the worker as a whole, together with other less interested skills. The fact that a worker's skills are not decomposable has made the labor economics problem difficult to solve.

Labor economics

An idea[8],[9] is to mathematically unbundle a worker's skills. The ideal(unrealistic) situation in which a firm can access to all skill components individually is an equilibrium. They show that at this equilibrium firms' wage distribution over employees is dominated in convex order by the distribution of workers' aggregated skills. In this context, testing of convex order between wage distribution and distribution of aggregated skills can be leveraged to test whether a labor market is at an equilibrium, and would clearly be valuable to government service when policies are being made.

[8] Choné, Kramarz, et al., *Matching Workers' Skills and Firms' Technologies: From Bundling to Unbundling*.

[9] Nordström Skans, Choné, and Kramarz, *When workers' skills become unbundled: Some empirical consequences for sorting and wages*.

Main goal

The goal of this talk is *How do we statistically and quantitatively test the convex order relation between two distributions, in an effective way?* Formally, we solve the following hypothesis problem:

$$\boldsymbol{H_0} : \mu \preceq \nu \quad \text{vs} \quad \boldsymbol{H_A} : \mu \not\preceq \nu. \tag{1}$$

Easy or hard?

If $\mu \not\preceq \nu$, $\sup_{\varphi\in\mathcal{A}} \left\{\int_{\mathbb{R}^d} \varphi d\mu - \int_{\mathbb{R}^d} \varphi d\nu\right\} = \infty$. Let μ_n and ν_n be empirical measures of μ and ν, respectively.

$$T_n := \sup_{\varphi\in\mathcal{A}} \left\{\int \varphi d\mu_n - \int \varphi d\nu_n\right\}.$$

On $\mathbb{R}^2$, $\mu = \delta_{(0,0)}$ and $\nu = \frac{1}{2}\delta_{(-1,0)} + \frac{1}{2}\delta_{(1,0)}$. It is easy check that $\mu \preceq \nu$. Also

$$\nu_n := \frac{1}{2}\delta_{(-1,0)} + \frac{1}{2}\delta_{\left(1,\frac{1}{n}\right)} \longrightarrow \nu \text{ weakly as } n \to \infty.$$

Clearly $\mu \not\preceq \nu_n$ for all $n \geq 1$. In fact, taking the set of convex functions which are zero on the line connecting $(-1, 0)$ and $\left(1, \frac{1}{n}\right)$, and arbitrarily large at $(0, 0)$ yields $T_n = \infty$.

Martingale optimal transport

Weak optimal transport

Gozlan, Roberto, Samson and Tetali[10] extends the classical optimal transport to the *weak optimal transport*. For $\pi \in \Pi(\mu, \nu)$, one can rewrite

$$\pi(dxdy) = p_x(dy)\mu(dx)$$

where $p_x(dy) = p(dy|x)$ be the probability kernel. Let $\theta : \mathbb{R}^d \times \mathcal{P}(\mathbb{R}^d) \to [0, \infty]$. The weak optimal transport cost $\mathcal{T}_\theta(\nu|\mu)$ is defined as

$$\mathcal{T}_\theta(\nu|\mu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \theta(x, p_x)\mu(dx).$$

If $\theta(x, p_x) = \int c(x, y)p_x(dy)$, then $\mathcal{T}_\theta(\nu|\mu)$ recovers the usual optimal transport.

[10] Gozlan et al., "Kantorovich duality for general transport costs and applications".

Martingale optimal transport

The *martingale optimal transport problem* is a variant of the weak optimal transport, which emerges from mathematics finance[11],[12].
$\mathcal{M}(\mu,\nu) \subset \Pi(\mu,\nu)$ is the subset of couplings satisfying $\mathbb{E}[Y|X] = X$ μ-a.e whenever $(X,Y) \sim \pi \in \mathcal{M}(\mu,\nu)$, i.e., the coupling π generates a martingale between μ and ν. The martingale optimal transport problem is defined as

$$\min_{\pi \in \mathcal{M}(\mu,\nu)} \int_{\mathbb{R}^d \times \mathbb{R}^d} c(x,y)\pi(dx,dy).$$

Recall that $\mathcal{M}(\mu,\nu)$ is non-empty if and only if $\mu \preceq \nu$[13].

[11] Hobson and Neuberger, "Robust bounds for forward start options".

[12] Beiglböck, Henry-Labordere, and Penkner, "Model-independent bounds for option prices—a mass transport approach".

[13] Strassen, "The Existence of Probability Measures with Given Marginals".

Relation

If one chooses

$$\theta(x, p_x) = \begin{cases} \int c(x, y) p_x(dy) & \text{if } \int y p_x(dy) = x, \\ \infty & \text{otherwise,} \end{cases}$$

then, the weak OT recovers the martingale OT.

Wasserstein projection

Convex order cone

Let

$$\theta(x, p_x) = \left(\int y p_x(dy) - x \right)^2,$$

and consider

$$T_2(\nu|\mu) := \inf_{\pi \in \Pi(\mu,\nu)} \int \left(\int y p_x(dy) - x \right)^2 \mu(dx).$$

Notice that $T_2(\nu|\mu) = 0$ if and only if $\mu \preceq \nu$ if and only if there is a martingale coupling between μ and ν.
Any connection to the classical OT?

Convex order cone

For $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, $\mathscr{P}^{cx}_{\preceq\nu}$ and $\mathscr{P}^{cx}_{\mu\preceq}$ are backward and forward convex order cones defined as

$$\mathscr{P}^{cx}_{\preceq\nu} := \left\{\xi \in \mathcal{P}_2(\mathbb{R}^d) : \xi \preceq \nu\right\}, \quad \mathscr{P}^{cx}_{\mu\preceq} := \left\{\eta \in \mathcal{P}_2(\mathbb{R}^d) : \nu \preceq \eta\right\}.$$

Kim and Ruan[14] discuss geodescially convexity of $\mathscr{P}^{cx}_{\preceq\nu}$ and $\mathscr{P}^{cx}_{\mu\preceq}$: only backward convex cone is geodescially convex.

[14] Kim and Ruan, "Backward and forward Wasserstein projections in stochastic order".

Wasserstein projection

Define projections onto $\mathscr{P}^{cx}_{\preceq\nu}$ and $\mathscr{P}^{cx}_{\mu\preceq}$ with respect to $\mathcal{W}_2$ distance:

$$\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right) := \inf_{\xi\in\mathscr{P}^{cx}_{\preceq\nu}} \mathcal{W}_2\left(\mu,\xi\right), \quad \mathcal{W}_2\left(\mathscr{P}^{cx}_{\mu\preceq}, \nu\right) := \inf_{\eta\in\mathscr{P}^{cx}_{\mu\preceq}} \mathcal{W}_2\left(\eta,\nu\right).$$

Theorem

(Gozlan and Juillet[a]) For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\mathcal{T}_2(\nu|\mu) = \mathcal{W}_2^2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right).$$

[a] Gozlan and Juillet, "On a mixture of Brenier and Strassen theorems".

Hence, $\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right) = 0$ if and only if $\mu \preceq \nu$.

Main questions

Recall the hypothesis problem:

$$\boldsymbol{H_0} : \mu \preceq \nu \quad \text{vs} \quad \boldsymbol{H_A} : \mu \not\preceq \nu.$$

Then, under $\boldsymbol{H_0} : \mu \preceq \nu$ we have $\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq \nu}\right) = 0$.
Let $\mu_n = \sum \frac{1}{n}\delta_{X_i}$ and $\nu_m = \sum \frac{1}{n}\delta_{Y_i}$ be empirical distributions of μ and ν, respectively. Our test statistics is $\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right)$, and the decision rule for (1) is

$$\text{Reject } \boldsymbol{H_0} \text{ if } \mathcal{W}_2(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}) \geq t(\alpha); \text{ Accept otherwise,} \qquad (2)$$

Is it really good enough for the hypothesis problem? If so, how can we compute it?

Stability of Wasserstein projection

Stability of Wasserstein projection

In order to be a good estimator, it should be consistent: as $n, m \to \infty$,

$$\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right) \to \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq \nu}\right).$$

In mathematical language, it should be stable.
In fact, Brückerhoff and Juillet[15] prove the instability of the martingale optimal transport problem in dimension $d \geq 2$.
But, we require the weakest stability: stability of the optimal value.

[15] Brückerhoff and Juillet, "Instability of martingale optimal transport in dimension $d \geq 2$".

Theorem

(Kim and Ruan[a]) For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, it holds that

$$\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq \nu}\right) = \mathcal{W}_2\left(\mathscr{P}^{cx}_{\mu \preceq}, \nu\right). \tag{3}$$

[a] Kim and Ruan, "Backward and forward Wasserstein projections in stochastic order".

Surprisingly helpful to prove the stability of $\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq \nu}\right)$.

Proof of Stability

Lemma

For any μ, ν, $\xi \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, it holds for Wasserstein backward projection that,

$$\left|\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right) - \mathcal{W}_2\left(\xi, \mathscr{P}^{cx}_{\preceq\nu}\right)\right| \leq \mathcal{W}_2\left(\mu, \xi\right). \tag{4}$$

Proof.

If $x, y \in X$ a metric space, and $C \subset X$ then the metric distance should satisfy $|d(x, C) - d(y, C)| \leq d(x, y)$; to see this notice

$$d(x, C) \leq d(x, y) + d(y, C),$$
$$d(y, C) \leq d(x, y) + d(x, C).$$

□

Proof of Stability

Lemma

For any μ, ν, $\xi \in \mathcal{P}_2\left(\mathbb{R}^d\right)$, it holds for Wasserstein backward projection that,

$$\left|\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\xi}\right)\right| \leq \mathcal{W}_2\left(\nu, \xi\right). \tag{5}$$

Proof.

$$\begin{aligned}
\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right) &= \mathcal{W}_2\left(\mathscr{P}^{cx}_{\mu\preceq}, \nu\right) \text{ by (3)} \\
&\leq \mathcal{W}_2\left(\mathscr{P}^{cx}_{\mu\preceq}, \xi\right) + \mathcal{W}_2\left(\xi, \nu\right) \text{ by (4)} \\
&= \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\xi}\right) + \mathcal{W}_2\left(\xi, \nu\right) \text{ by (3).}
\end{aligned}$$

Similarly,

$$\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\xi}\right) \leq \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right) + \mathcal{W}_2\left(\nu, \xi\right).$$

Proof of Stability

Theorem (Quantitative stability)

For any probability measures μ, μ' and ν, ν',

$$\left|\mathcal{W}_2\left(\mu', \mathscr{P}^{cx}_{\preceq\nu'}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right)\right| \leq \mathcal{W}_2\left(\mu, \mu'\right) + \mathcal{W}_2\left(\nu, \nu'\right). \quad (6)$$

Proof.

Observe that

$$\begin{aligned}
&\left|\mathcal{W}_2\left(\mu', \mathscr{P}^{cx}_{\preceq\nu'}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right)\right| \\
&\leq \left|\mathcal{W}_2\left(\mu', \mathscr{P}^{cx}_{\preceq\nu'}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu'}\right) + \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu'}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right)\right| \\
&\leq \left|\mathcal{W}_2\left(\mu', \mathscr{P}^{cx}_{\preceq\nu'}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu'}\right)\right| + \left|\mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu'}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq\nu}\right)\right|.
\end{aligned}$$

Applying (4) and (5) for the first term and the second term in the above last line, respectively, the conclusion follows. □

Consistency

Corollary (Consistency of $\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right)$)

Let μ_n and ν_m be empirical distributions drawn from μ and ν, respectively. Then,

$$\left|\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right) - \mathcal{W}_2\left(\mu, \mathscr{P}^{cx}_{\preceq \nu}\right)\right| \leq \mathcal{W}_2\left(\mu, \mu_n\right) + \mathcal{W}_2\left(\nu, \nu_m\right). \quad (7)$$

So, our estimator converges to the true one.

Converegence rate

The convergence rate of $\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right)$ can be derived by combining with many recent progresses of the convergence rate of empirical distributions under some mild assumptions (we leverage Fournier and Guillin[16], Weed and Bach[17] and Lei[18]).

[16] Fournier and Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure".

[17] Weed and Bach, "Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance".

[18] Lei, "Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces".

Theorem (Convergence rate of $\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right)$)

(i) Assume that μ and ν satisfy the log-Sobolev inequality with a constant $\kappa > 0$. Then, for all sufficiently large n, m,

$$\left|\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right) - \mathcal{W}_2(\mu, \mathscr{P}^{cx}_{\preceq \nu})\right| \leq O\left((n \wedge m)^{-\left(\frac{1}{d} \wedge \frac{1}{4}\right)} (\log(n \wedge m))^{\frac{1}{2}\mathbb{1}_{d=4}}\right)$$

with probability $1 - 2\exp\left(-\frac{\sqrt{n \wedge m}}{2\kappa}\right)$.

(ii) Assume that μ and ν have bounded supports with diameter at most D. Let $d_p^*(\mu)$ be the Wasserstein dimension of μ. If $k > d_2^*(\mu) \vee d_2^*(\nu) \vee 4$, then for all sufficiently large n, m,

$$\left|\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right) - \mathcal{W}_2(\mu, \mathscr{P}^{cx}_{\preceq \nu})\right| \leq O\left((n \wedge m)^{-\frac{1}{k}}\right)$$

with probability $1 - 2\exp\left(\frac{-2\sqrt{n \wedge m}}{D^4}\right)$.

Computation scheme

How to compute

A next question is how to compute $\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_m}\right)$. For this purpose, we need the following important theorem.

Theorem

(Gozlan and Juillet[a], Kim and Ruan[b]) For any $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$, there exists a unique (backward) projection of μ, denoted by $\overline{\mu}$, onto $\mathscr{P}^{cx}_{\preceq \nu}$. Furthermore,

$$\overline{\mu} = (\nabla \varphi)_{\#} \mu, \tag{8}$$

where φ is a proper lower semicontinuous convex function such that $D^2\varphi \leq Id$.

[a] Gozlan and Juillet, "On a mixture of Brenier and Strassen theorems".
[b] Kim and Ruan, "Backward and forward Wasserstein projections in stochastic order".

This is very similar to Breinier's theorem: but it does not require the absolute continuity.

Simple case

Theorem

(Gozlan and Juillet[a]) Let $\{y_1, \ldots, y_k\}$ for $k \leq d$ be the set of vertices of a simplex and ν_k be an atomic measure supported on it. Denote Δ as the convex hull of $\{y_1, \ldots, y_k\}$. Then, for any $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, there is $v \in \mathbb{R}^d$ such that the map $T : \mathbb{R}^d \to \mathbb{R}^d$ defined by

$$T(x) = proj_{\Delta}(x + v)$$

is such that $\overline{\mu} = (T_{\#})(\mu)$.

[a] Gozlan and Juillet, "On a mixture of Brenier and Strassen theorems".

The idea is to use martingale property: since $\overline{\mu} \preceq \nu$, the mean(barycenter) should be equal. Also, spt$(\overline{\mu}) \subseteq \Delta$. In fact, the barycenter and the geometry of spt$(\overline{\mu})$(lying in the convex hull of spt(ν_m)) almost determine $\overline{\mu}$.

Finite case

For an arbitrary finite set, the argument is insufficient. Consider $\mu = \frac{1}{2}\delta_{e_1} + \frac{1}{2}\delta_{-e_1}$ and $\nu = \frac{1}{2}\delta_{e_2} + \frac{1}{2}\delta_{-e_2}$ for the standard basis vectors e_1, e_2 of $\mathbb{R}^2$. These μ, ν have the same barycenter, but, there is no convex order between them.

It is because of the mass transport constraint: there are different extreme points of given convex hull having the same barycenter. We should use all points to write the barycenter, i.e., each point of $\mathrm{spt}(\nu_m)$ should receive some mass from μ.

Finite case

Lemma

For $m \in \mathbb{N}$, let Δ_m be the m-simplex. Assume that ν_m is a distribution over m-points, $\{y_1, \ldots, y_m\}$. Let $T : \Delta_m \longrightarrow conv(\text{spt}(\nu_m))$ such that $T(\boldsymbol{\alpha}) = \sum_{j=1}^{m} \alpha_j y_j$ for each $\boldsymbol{\alpha} := (\alpha_1, \ldots, \alpha_m)^T \in \Delta_m$. Then,

$$\mathscr{P}^{cx}_{\preceq \nu_m} = \left\{ (T)_{\#}(\omega) : \omega \in \mathcal{P}(\Delta_m) \text{ s.t. } \int_{\Delta_m} \alpha_j d\omega(\boldsymbol{\alpha}) = \nu_m(y_j) \right\}.$$

Computing scheme

Theorem

For each $x_i \in \text{spt}(\mu_n)$ *we write* $\boldsymbol{\alpha}(x_i) := (\alpha_1(x_i), \ldots, \alpha_m(x_i))^T$. *Consider the following constrained minimization problem:*

$$\min_{\{\boldsymbol{\alpha}(x_1),\ldots,\boldsymbol{\alpha}(x_n)\}} \sum_{i=1}^{n} \mu_n(x_i) \left(\sum_{j=1}^{m} \alpha_j(x_i) y_j - x_i \right)^2$$

$$s.t.\ \boldsymbol{\alpha}(x_i) \in \Delta_m, \quad \sum_{i=1}^{n} \alpha_j(x_i) \mu_n(x_i) = \nu_m(y_j).$$

Then, there exists a unique minimizer $\{\boldsymbol{\alpha}^*(x_1), \ldots, \boldsymbol{\alpha}^*(x_n)\}$ *which is the projection of* μ_n *onto* $\mathscr{P}^{cx}_{\preceq \nu_m}$, *denoted by* $\overline{\mu}_n$, *given in this way: for any measurable* $E \subseteq conv(\text{spt}(\nu_m))$,

$$\overline{\mu}_n(E) = \sum_{i=1}^{n} \mu_n(x_i) \mathbb{1}_E \left(\sum_{j=1}^{m} \alpha_j^*(x_i) y_j \right).$$

Computing scheme

Let $x_i = (x_{i1}, \cdots, x_{id})$, $y_j = (y_{j1}, \cdots, y_{jd})$ and

$$\mathsf{A} = \begin{bmatrix} \alpha_1(x_1) & \dots & \alpha_m(x_1) \\ \vdots & \ddots & \vdots \\ \alpha_1(x_n) & \dots & \alpha_m(x_m) \end{bmatrix}, \quad \mathsf{Y} = \begin{bmatrix} y_{11} & \dots & y_{1d} \\ \vdots & \ddots & \vdots \\ y_{m1} & \dots & y_{md} \end{bmatrix}$$

$$\mathsf{X} = \begin{bmatrix} x_{11} & \dots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nd} \end{bmatrix}, \quad \boldsymbol{\mu_n} = \begin{bmatrix} \mu_n(x_1) \\ \vdots \\ \mu_n(x_n) \end{bmatrix}, \quad \boldsymbol{\nu_m} = \begin{bmatrix} \nu_m(y_1) \\ \vdots \\ \nu_m(y_m) \end{bmatrix}.$$

It can be written as

$$\min_{\mathsf{A}} \operatorname{trace}\left((\mathsf{AY} - \mathsf{X})^T \operatorname{diag}(\boldsymbol{\mu_n})(\mathsf{AY} - \mathsf{X})\right)$$
$$\text{s.t. } \mathsf{A}^T \boldsymbol{\mu_n} = \boldsymbol{\nu_n}, \quad \mathsf{A}1_{\boldsymbol{m}} = 1_{\boldsymbol{n}}, \quad \mathsf{A} \geq 0.$$

Experiment

H_0 case

On $\mathbb{R}^2$, let $N(0, I_2)$ be the standard Gaussian distribution. Consider $\mu = \text{Unif}\,[0,1]^2$ and $\nu = \text{Unif}\,[0,1]^2 * N(0, I_2)$. $\mu \preceq \nu$ since for $X \sim \mu$ and $Y \sim \nu$, the martingale condition $\mathbb{E}_\pi[Y|X] = X$ is achieved by $\pi(y|x) = N(x, I_2)$, the isotropic Gaussian with mean x. Let μ_n and ν_n be the empirical distributions of μ and ν with independent n-samples, respectively. Also, both have the uniform weight over samples. In this experiment, we use CVXPY python code developed by Diamond and Boyd[19] and Agrawal et al.[20].

[19] Diamond and Boyd, "CVXPY: A Python-embedded modeling language for convex optimization".

[20] Agrawal et al., "A rewriting system for convex optimization problems".

H_0 case

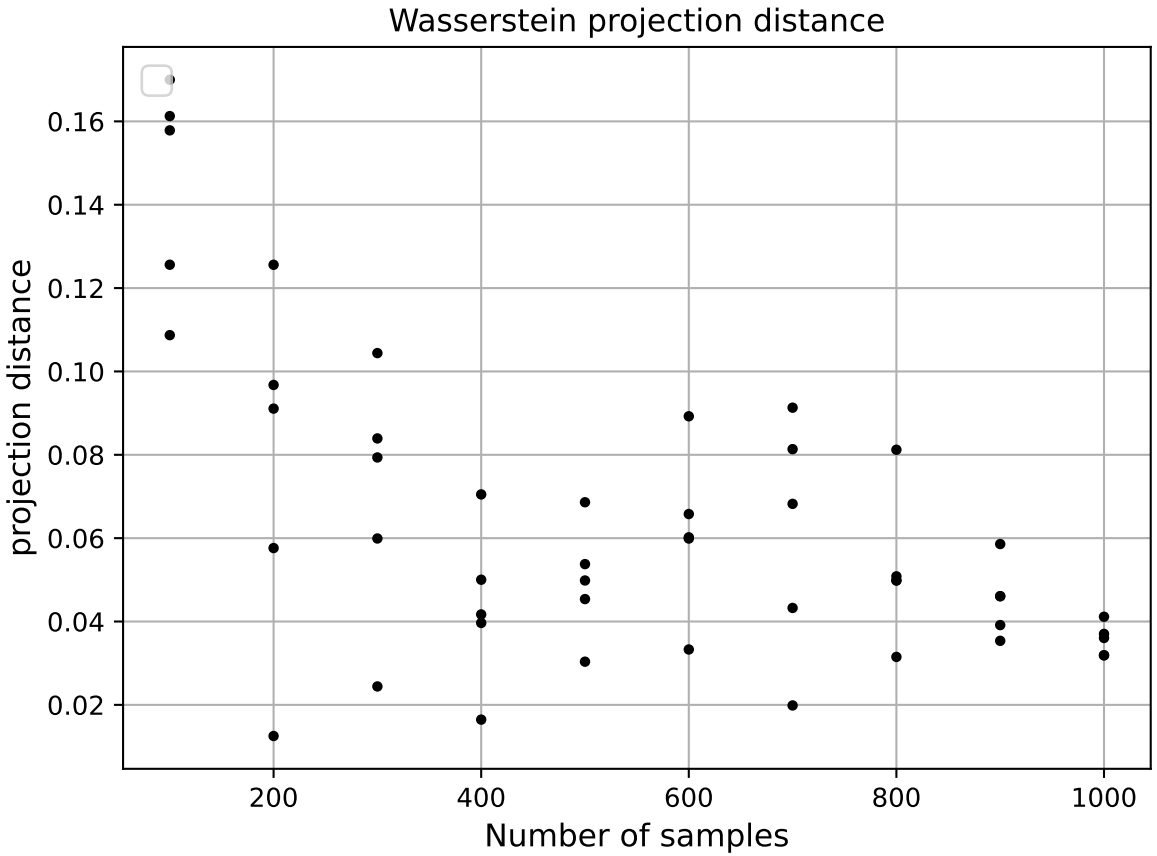


Figure: The plot of $\mathcal{W}_2\left(\mu_n, \mathscr{P}^{cx}_{\preceq \nu_n}\right)$.

H_0 case

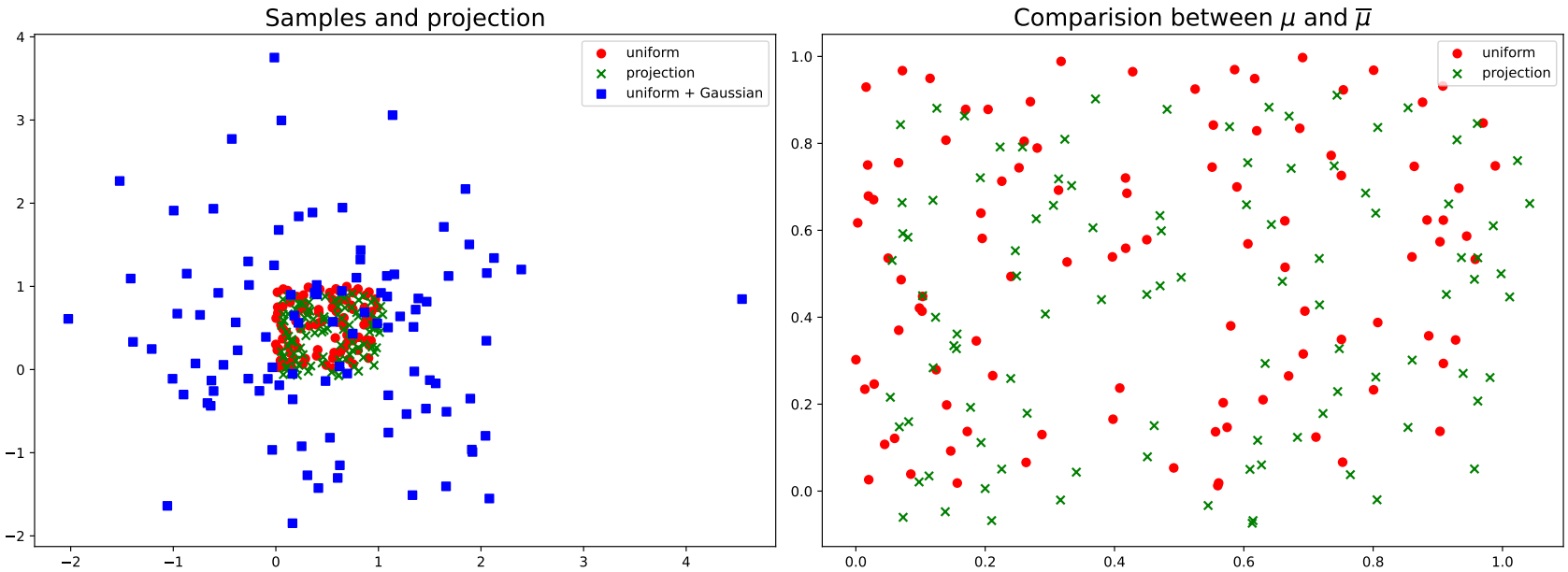


Figure: The geometry of Wasserstein projection.

Entropic Frank-Wolfe Algorithm

Let

$$\min_{\substack{\pi\in\mathbb{R}^n\times\mathbb{R}^n \\ \pi 1=1,\pi^T 1=1}} \mathcal{J}(\pi) \triangleq \frac{1}{n}\|\pi \mathsf{Y}-\mathsf{X}\|_{\mathsf{F}}^2 .$$

Taking the gradient descent type algorithm:

$$\min_{\pi\in\mathcal{E}} \nabla\mathcal{J}(\pi_k)\odot\pi,$$

where $\odot$ is the Hadamard product (the element-wise product) of matrices,

$$\nabla\mathcal{J}(\pi_k) = \nabla_\pi\left(\frac{1}{n}\|\pi \mathsf{Y}-\mathsf{X}\|_{\mathsf{F}}^2\right)\Big|_{\pi_k} = \frac{2}{n}(\pi_k \mathsf{Y}-\mathsf{X})\,\mathsf{Y}^T$$

and

$$\mathcal{E} = \left\{\pi\in\mathbb{R}^n\times\mathbb{R}^n : \pi 1 = 1, \pi^T 1 = 1\right\}.$$

Entropic Frank-Wolfe Algorithm

Consider

$$\min_{\gamma \in \mathcal{E}} \left\{ \nabla \mathcal{J}\left(\pi_k\right) \odot \gamma + \varepsilon_k \mathsf{KL}\left(\gamma \,|1 \otimes 1\right) \right\}.$$

ε_k satisfies $\varepsilon_k \to 0$ as $k \to \infty$. Note if we set $\varepsilon_k = 0$, then we see that the entropic version recovers the previous one.

Consider two dimensional distributions

$$\mu \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -2 \\ -2 & 3 \end{pmatrix}\right), \ \nu \sim N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}\right).$$

Clearly $\mu \not\preceq \nu$. Now 10^4 samples are drawn respectively from μ and ν. The empirical measures are written as μ_n and ν_n. We start the entropic Frank-Wolfe algorithm with initial π_0 being a $10^4 \times 10^4$ matrix with each row equal to

$$\left(10^{-4}, ..., 10^{-4}\right) \in \mathbb{R}^{10^4}.$$

This corresponds to a Dirac measure on the cone $\mathscr{P}^{\mathsf{cx}}_{\preceq \nu_n}$ concentrating on the barycenter of ν_n. In around 10 iterations, the algorithm reaches a minimal value close to 2 and stabilizes in further iterations, which indicates μ_n is not in the cone $\mathscr{P}^{\mathsf{cx}}_{\preceq \nu_n}$. Figure 4 shows some snapshots of the evolution of the probability along the cone.

Entropic Frank-Wolfe Algorithm

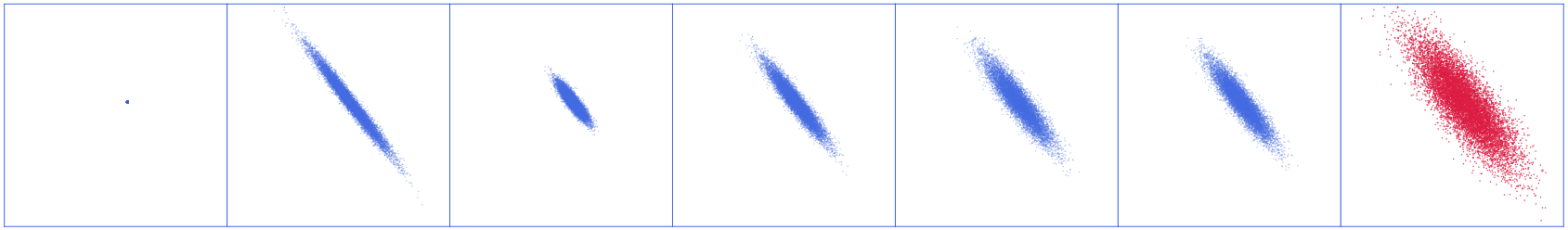


Figure: Left-most: the initial distribution concentrating on the barycenter of ν_n. Right-most: the empirical distribution ν_n. Middle: slices of the gradient flow along the convex order cone $\mathscr{P}^{\mathsf{cx}}_{\preceq \nu_n}$ (from left to right) generated by the entropic Frank-Wolfe algorithm.

Conclusions and future works

Conclusions

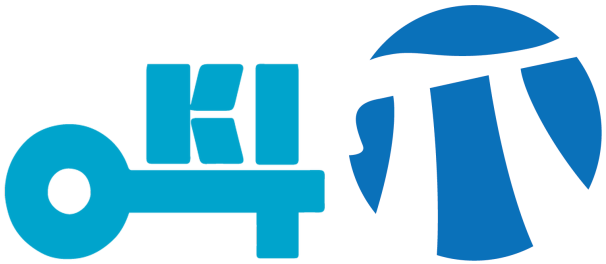
In this work

- A new test statistic for the hypothesis problem of convex order
- The stability of the projected Wasserstein distance, and as a byproduct of it the consistency of the test statistic
- The rate of convergence under mild assumptions
- Computation scheme and algorithm
- Experiment with synthetic data

Future works

- Other stochastic order?
- Sharpen the theory: e.g. central limit theorem
- Boosting the rate of convegence by entorpic optimal transport
- Faster algorithm

Thank you for your attention!



Pacific Institute *for the* Mathematical Sciences