Distributionally robust sample average approximation

Andy Philpott
Engineering Science
University of Auckland

(based on joint work with Eddie Anderson and Dominic Keehan)

ASL60 Workshop, Seattle, August 9-10, 2022.

Infinite transportation problems

Scholar

[CITATION] **Extreme points of infinite transportation problems**

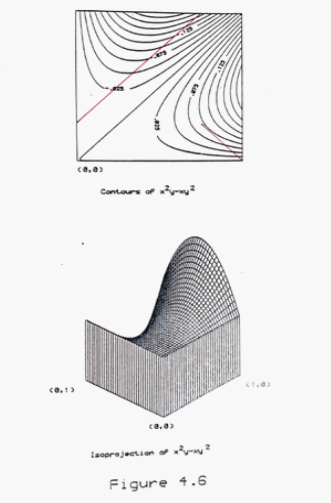
A Lewis - Methods Operating Research, 1986

Save Cite Cited by 3 Related articles

Showing the best result for this search. See all results

$$
\begin{array}{lll}
\min & \int_{\mathcal{N}} \int_{\mathcal{M}} c(u,v)\rho(u,v)dxdy & \\
\text{s.t.} & \int_{\mathcal{N}} \rho(u,v)du = \mu(u), & u \in \mathcal{M}, \\
 & \int_{\mathcal{M}} \rho(u,v)dv = \nu(v), & v \in \mathcal{N}, \\
 & \rho(u,v) \geq 0, & (u,v) \in \mathcal{M} \times \mathcal{N}.
\end{array}
$$

A picture from my PhD thesis



Transport $U(0,1)$ mass to $U(0,1)$ when $c(u,v) = uv(u-v)$

Gaspard Monge and Leonid Kantorovich

Source: Wikipedia

Koopmans, Dantzig, Kantorovich (Laxenburg, 1976)

Source: "History of mathematical programming", Lenstra, Rinnooy-Kan, Schrijver.

Wasserstein distance (a.k.a. Kantorovich Metric)

Suppose $(\mathcal{M}, d)$ is a Polish metric space. For $p \geq 1$ let $\mathcal{P}_p(\mathcal{M})$ denote the collection of all probability measures μ on M with finite pth moment. For distributions $\mathbb{P}, \mathbb{Q} \in \mathcal{P}_p(\mathcal{M})$ let $\Gamma(\mathbb{P}, \mathbb{Q})$ denote the set of joint distributions with marginals $\mathbb{P}$ and $\mathbb{Q}$. For $p \geq 1$ the pth Wasserstein distance under the metric d is defined as:

$$W^p(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{M} \times \mathcal{M}} d(u, v)^p d\gamma(u, v) \right)^{\frac{1}{p}} .$$

Example: total variation

$\mathcal{M} = \{z_1, z_2, z_3, \ldots, z_N\}$, and d is the discrete metric

$$d(u,v) = \begin{cases} 0, & u = v \\ 1, & \text{otherwise} \end{cases}.$$

In this case $\mathbb{P}$ and $\mathbb{Q}$ can be represented by probability distributions p, q supported on $\mathcal{M}$, and $W^1(\mathbb{P}, \mathbb{Q})$ reduces to the total-variation distance $\sum_{i=1}^{N} |q_i - p_i|$.

Examples

When $\mathcal{M} \subseteq \mathbb{R}^n$ and $d(u, v) = \|u - v\|$ (the standard Euclidean norm) then

$$W^1(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{M} \times \mathcal{M}} \|u - v\| d\gamma(u, v),$$

$$W^2(\mathbb{P}, \mathbb{Q}) = \left(\inf_{\gamma \in \Gamma(\mathbb{P}, \mathbb{Q})} \int_{\mathcal{M} \times \mathcal{M}} \|u - v\|^2 d\gamma(u, v) \right)^{\frac{1}{2}}.$$

Stochastic optimization

- Stochastic optimization problem with random cost function $c(x, Z)$

$$\text{SO: } \min_x \mathbb{E}_{\mathbb{P}}[c(x, Z)].$$

- Expectations are taken over the random variable Z, with instance $z \in \mathbb{R}^m$, and probability distribution $\mathbb{P}$.
- We call SO the true problem.

Definition

Given any x, $\bar{c}(x) = \mathbb{E}_{\mathbb{P}}[c(x, Z)]$ is the out-of-sample cost of x evaluated with the true probability distribution $\mathbb{P}$.

- Optimal solution of SO is denoted x^*, optimal objective value $C^* = \bar{c}(x^*) = \mathbb{E}_{\mathbb{P}}[c(x^*, Z)]$.

Resrict attention to convex quadratic programs

$$\text{SO:}\ \min_x \mathbb{E}_{\mathbb{P}}[c(x,Z)]$$

Let $c(x,z) = \frac{1}{2}x^\top H(z)x - v(z)^\top x$ where $H(z)$ is positive definite a.s.

$$\bar{c}(x) = \frac{1}{2}x^\top \mathbb{E}_{\mathbb{P}}[H]x - \mathbb{E}_{\mathbb{P}}[v]^\top x.$$

- Optimal solution $x^* = \mathbb{E}_{\mathbb{P}}[H]^{-1}\mathbb{E}_{\mathbb{P}}[v]$
- Optimal objective function value is

$$\bar{c}(x^*) = -\frac{1}{2}\mathbb{E}_{\mathbb{P}}[v]^\top \mathbb{E}_{\mathbb{P}}[H]^{-1}\mathbb{E}_{\mathbb{P}}[v].$$

Sample average approximation

- The decision maker does not know $\mathbb{P}$, but has a sample, $S = \{z_1, z_2, ..., z_N\}$, of Z.
- Write $\mathbb{P}_0$ for the sample distribution which has probability $\frac{1}{N}$ at each of the sample points in $S = \{z_1, z_2, ..., z_N\}$.
- Approximate the value of $\mathbb{E}_{\mathbb{P}}[c(x, Z)]$ by $\mathbb{E}_{\mathbb{P}_0}[c(x, Z)]$ and solve the sample average approximation problem

$$\text{SAA:} \quad \min_{x \in X} \mathbb{E}_{\mathbb{P}_0}[c(x, Z)].$$

- Let $x_0(S)$ denote the solution of SAA (depends on sample S).
- $\bar{c}(x_0(S)) = \mathbb{E}_{\mathbb{P}}[c(x_0(S), Z)]$ is the out-of-sample cost of $x_0(S)$ evaluated with the true probability distribution $\mathbb{P}$. Note: this depends on sample S.

Post decision disappointment

- SAA solution value is biased low

$$\mathbb{E}_{\mathbb{S}}\left[\mathbb{E}_{\mathbb{P}_0}[c(x_0(S), Z)]\right] \leq \bar{c}(x^*)$$

- Out of sample cost of $x_0(S)$ is never lower than $\bar{c}(x^*)$, so

$$\bar{c}(x^*) \leq \bar{c}(x_0(S))$$

- Promise of SAA solution is not delivered when evaluated out of sample.
- Consider some robustification $x_\delta(S)$ of SAA solution.

Definition

Value of robustification, $\mathrm{VRS}(\delta) = \mathbb{E}_S[\bar{c}(x_0(S)) - \bar{c}(x_\delta(S))]$.

- When is $\mathrm{VRS}(\delta) > 0$?

Distributionally robust optimization (DRO)

[Scarf, 1958, Zackova, 1966, Pflug and Wozabal, 2007, ...]

- Distributionally robust optimization (DRO) solves the following problem
$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}}\left[c(x, Z)\right],$$
for some choice of $\mathcal{P}_\delta$ being a ball of size δ centered at $\mathbb{P}_0$.
- We write $x_\delta(S)$ for the optimal solution of DRO and write $C_\delta(S) = \bar{c}(x_\delta(S))$ (the out-of-sample cost of $x_\delta(S)$).
- When $\delta = 0$ we have $\mathcal{P}_\delta = \{\mathbb{P}_0\}$, so then $C_\delta(S) = \bar{c}(x_0(S))$.
- We define $\mathcal{P}_\delta$ for $\delta > 0$ using a Wasserstein metric.
- If δ chosen large enough then true distribution $\mathbb{P}$ lies in $\mathcal{P}_\delta$ with high probability (Fournier & Guillin, 2015). So, with high probability,
$$\mathbb{E}_{\mathbb{Q}^*}\left[c(x_\delta(S), Z)\right] \geq \mathbb{E}_{\mathbb{P}}\left[c(x_\delta(S), Z)\right] = \bar{c}(x_\delta(S)).$$

Out-of-sample performance for quadratics

Suppose $C(x, z) = \frac{1}{2}x^\top H(z)x^2 - v(z)^\top x$. Let $\bar{x}_\delta = \mathbb{E}_S[x_\delta(S)]$.

Definition

The cost bias of $x_\delta(S)$ is

$$\beta_\delta = \bar{c}(\bar{x}_\delta) - \bar{c}(x^*).$$

Definition

The variation of $x_\delta(S)$ is

$$V_\delta = \mathbb{E}_S[(x_\delta(S) - \bar{x}_\delta)^\top \mathbb{E}[H](x_\delta(S) - \bar{x}_\delta)]$$

Proposition

If $C(x, z) = \frac{1}{2}x^\top H(z)x^2 - v(z)^\top x$ *then*

$$VRS(\delta) = \frac{1}{2}(V_0 - V_\delta) - (\beta_\delta - \beta_0).$$

Examples

Example (1)

$c(x, z) = \frac{1}{2}(x - z)^2$

Example (2)

$c(x, z) = \frac{1}{2}x^2 - g(z)x$

Example 1 with W1

$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}} \left[\frac{1}{2}(x - Z)^2 \right],$$

$$\mathcal{P}_\delta = \{\mathbb{Q} \in \mathcal{P}_1(\mathcal{M}) : \inf_{\gamma \in \Gamma(\mathbb{P}_0, \mathbb{Q})} \int_{\mathcal{M} \times \mathcal{M}} \|u - v\| d\gamma(u, v) \leq \delta\}$$

If $\mathcal{M} = (-\infty, \infty)$ then supremum not attained: $\mathbb{Q}$ sends atoms to infinity.
If $\mathcal{M} = [a, b]$, then supremum attained: $\mathbb{Q}$ sends mass to a or b. For $\delta > b - a$, solution to DRO is $x_\delta(S) = \frac{a+b}{2}$.

Example 2 with W1

$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}} \left[c(x, Z) \right],$$

$$\mathcal{P}_\delta = \{ \mathbb{Q} \in \mathcal{P}_1(\mathcal{M}) : \inf_{\gamma \in \Gamma(\mathbb{P}_0, \mathbb{Q})} \int_{\mathcal{M} \times \mathcal{M}} \|u - v\| d\gamma(u, v) \leq \delta \}$$

Proposition (Anderson and P., 2021)

When $c(x, z) = \frac{1}{2}x^2 - g(z)x$ and g is a strictly convex and non-negative function of z then $VRS(\delta) > 0$ when

$$\mathbb{E}_S \left[(\bar{g}_0(S) - \bar{g}) \|\nabla g(z^*(S))\| \right] > 0$$

where $z^*(S) = \arg\max_{z_i \in S} \{ \|\nabla g(z_i)\| \}$.

Example 1 with W2

$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}} \left[(x - Z)^2 \right]$$

$$\mathcal{P}_\delta = \{\mathbb{Q} \in \mathcal{P}_2(\mathcal{M}) : \inf_{\gamma \in \Gamma(\mathbb{P}_0, \mathbb{Q})} \int_{\mathcal{M} \times \mathcal{M}} \|u - v\|^2 d\gamma(u, v) \le \delta^2\}$$

Suppose $\mathcal{M} = (-\infty, \infty)$. Then

$$x_\delta(S) = \frac{1}{N} \sum_{i=1}^{N} z_i.$$

Thus $\beta_\delta = \beta_0$, $V_\delta = V_0$ and $\mathrm{VRS}(\delta) = 0$.

Example 2 with total variation

$$\text{DRO:} \quad \min_{x \in X} \sup_{\mathbb{Q} \in \mathcal{P}_\delta} \mathbb{E}_{\mathbb{Q}} \left[\frac{1}{2} x^2 - Zx \right]$$

Given a sample $S = \{z_1, z_2, ..., z_N\}$

$$\mathcal{P}_\delta = \{(q_1, q_2, ..., q_N) : \sum_{i=1}^{N} |q_i - p_i| \leq \delta\}.$$

$$\begin{aligned} \text{DRO:} \min_x \quad & \max_q \textstyle\sum_i q_i (\frac{1}{2} x^2 - z_i x) \\ & \textstyle\sum_{i=1}^N \left| q_i - \frac{1}{N} \right| < \delta, \\ & \textstyle\sum_{i=1}^N q_i = 1, \quad q \geq 0. \end{aligned}$$

Example 2 with total variation

Proposition (Anderson and P., 2021)

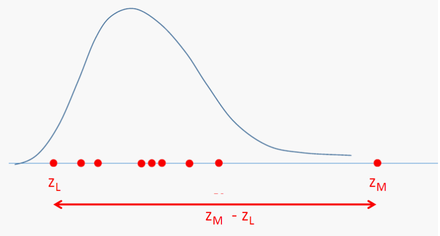
Suppose $c(x,z) = \frac{1}{2}x^2 - zx$ *with* $z > 0$ *almost surely, and sample* $S = \{z_1, z_2, ..., z_N\}$ *and* $z_0(S) = \frac{1}{N}\sum_{i=1}^{N} z_i$*. Then*

$$VRS(\delta) = (\delta/2)cov(z_0(S), R(S)) - \left(\delta^2/8\right)\mathbb{E}_S[R(S)^2],$$

where $R(S) = z_M - z_L$*. If the distribution of* Z *is symmetric about its mean then the* $(\delta/2)$ *term is zero and* $VRS(\delta) < 0$ *for all* δ*. If* $cov(z_0(S), R(S)) > 0$ *then* $VRS(\delta) > 0$ *for small* δ*.*

Total variation improvement with right skew

- If F has a large right tail, $z_0(S)$ and $z_M - z_L$ are both large for samples with $z_M \ggg 0$



- This implies that $z_0(S)$ and $z_M - z_L$ are positively correlated, so $\text{cov}(z_0(S), R(S)) > 0$.
- Robustifying takes weight from high-price outlier z_M and moves it to z_L, giving $\text{VRS}(\delta) > 0$.

Example 1 with total variation

$$\text{SO:} \; \min_x \mathbb{E}_{\mathbb{P}}[\frac{1}{2}(x - Z)^2]$$

Proposition

Suppose sample $S = \{z_1, z_2, ..., z_N\}$ *with order statistics* z_L *and* z_M. *If* $\delta > 1 - \frac{2}{N}$ *then* $\beta_\delta = \beta_0 = 0$, *and*

$$VRS(\delta) = \frac{1}{2}\left(\frac{\sigma^2}{N} - \mathbb{E}_S[(\frac{z_L + z_M}{2} - \mu)^2]\right)$$

Corollary

If $\delta > 1 - \frac{2}{N}$ *then* $VRS(\delta) > 0$ *for uniform* Z.

Ronald Fisher 1922

Suppose one takes 100 samples of a $U(0,1)$ random variable, and orders the sample so

$$z_1 \leq z_2 \leq \ldots \leq z_{100}.$$

The variance of the sample average is $\frac{\sigma^2}{N} \approx 8.33 \times 10^{-4}$. The variance of the first order statistic z_1 (and of the Nth order statistic z_{100}) is $\frac{N}{(N+1)^2(N+2)} \approx 10^{-4}$. So (assuming z_1 and z_{100} are independent) the variance of $\frac{u_1+u_{100}}{2} \approx \frac{10^{-4}}{2} \approx \frac{\sigma^2}{16N}$

$$\mathrm{VRS}(\delta) \approx \frac{15\sigma^2}{32N} > 0$$

(Source: Wikipedia)

Happy Birthday!

Clare College, March 17, 2019

References

- Anderson, E.J. and Philpott, A.B., Improving sample average approximation using distributional robustness, *INFORMS Journal on Optimization*, 2021.
- Esfahani, P.M. and Kuhn, D., 2017. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, pp.1-52.
- Fisher, R.A., 1922. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London*. Series A, 222(594-604), pp.309-368.
- Fournier, N. and Guillin, A., 2015. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3), pp.707-738.
- Gotoh, J.Y., Kim, M.J. and Lim, A.E., Calibration of distributionally robust empirical optimization models. arXiv preprint arXiv:1711.06565, 2017.

References

- Gotoh, J.Y., Kim, M.J. and Lim, A.E., Robust empirical optimization is almost the same as mean–variance optimization. *Operations Research Letters*, 2018.
- Kuhn, D., Esfahani, P.M., Nguyen, V.A. and Shafieezadeh-Abadeh, S., 2019. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In Operations research & management science in the age of analytics (pp. 130-166). INFORMS.
- Pflug, G.C., 2001. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming*, 89(2), pp.251-271.
- Pflug, G. and Wozabal, D., Ambiguity in portfolio selection. *Quantitative Finance* 7(4) 435-442, 2007.
- Scarf, H., A min-max solution of an inventory problem, in *Studies in the Mathematical Theory of Inventory and Production*, 201-9, 1958.
- Žáčková, J., On minimax solutions of stochastic linear programming problems. *Časopis pro pěstování matematiky*, 91(4), 423-430, 1966.