Wasserstein metrics and empirical distributions in stability of stochastic programs[1]

Michal Houda
Charles University of Prague, Faculty of Mathematics and Physics, Czech Republic

Abstract

Practical economic problems often ask for optimization procedures, not unfrequently with random inputs leading thus to stochastic programming models. The randomness is modelled through the underlying probability distribution, which is assumed to be completely known. But the ”true“ probability distribution is rarely fully known; instead, approximation or estimates are used. Some kind of stability of optimal values and optimal solution sets with respect to changes in the probability distribution is then required. This paper illustrates how a particular distance — Wasserstein metrics, measuring a distance between two probability distribution — affects the stability of stochastic optimization problems. Numerical examples for a choice of distributions and their empirical estimates are given.

1 Introduction

Stochastic programming specializes in problems where the uncertainty of input parameters has to be taken into account. Usually, the randomness is introduced to the model via probability distribution of the random variable. In such models, a full knowledge of the distribution is required. But in practice, its estimates and approximations have to be applied instead, due to modelling and numerical difficulties; for example, if the distribution for some input random parameter can be only estimated from a historical data series.

In a stochastic programming model, when replacing the original distribution with its approximation or estimate, one have to be cautious about resulting changes in optimal value and/or in optimal solution to the problem. We speaks about the *stability analysis* of stochastic programs *with respect to changes in* the underlying *probability measure.* Apart from other things, the stability analysis introduce another non-trivial task – to choose a suitable distance on the space of probability measures. A great attention has already been paid to this area in the literature, see e. g. [1], [2], [4], [5], [12], [13], etc. It turns out again that the problem is closely related to the properties of the original model and cannot be treated separately from it.

Refering to recent results in the stability of stochastic programs (see e. g. [10]), we consider the Wasserstein metrics in the remainder of the paper — it is a convenient distance in many cases (see also Section 3). Let $\mathcal{P}(\Xi)$ be the space of all Borel probability measures on some

[1]This research was supported by the Grant Agency of the Czech Republic under Grant 402/01/0539

Borel set $\Xi \subset \mathbb{R}^s$ and denote

$$\mathcal{P}_1(\Xi) := \left\{ \nu \in \mathcal{P}(\Xi) : \int_\Xi \|\xi\| \nu(\mathrm{d}\xi) < +\infty \right\}$$

the set of probability measures having finite first moment. Let $\mu, \nu \in \mathcal{P}_1(\Xi)$. The *1-Wasserstein metrics* is then defined by

$$W_1(\mu, \nu) := \inf_{\eta \in D(\mu,\nu)} \int_{\Xi \times \Xi} \|\xi - \tilde{\xi}\| \, \eta(\mathrm{d}\xi \times \mathrm{d}\tilde{\xi}),$$

where $D(\mu, \nu)$ is the set of all probability measures (of $\mathcal{P}(\Xi \times \Xi)$), for which μ and ν are marginal distributions. The practical dimension of the metrics appears when dealing with one-dimensional random variables; then the Wasserstein metrics reads (see [15])

$$W_1(\mu, \nu) = \int_{-\infty}^{+\infty} |F(t) - G(t)| \mathrm{d}t \tag{1}$$

where F and G are distribution functions corresponding to the probability measures μ and ν. In the latter case, the metrics coincides with the Fortet-Mourier metrics, closely related to a more general concept of distances having ζ-structure. For details, we refer to the papers [1], [10], and the book [9]. The main disability of the Wasserstein distance is recognized while dealing with distributions having heavy tails (see [6]).

At the end of the paper, we consider the empirical distribution as the selected approximation method. The one-dimensional *empirical distribution function*, based on the sample of i. i. d. random variables $\xi_1, \xi_2, \ldots$ with common distribution function F, is defined by

$$F_n(z) = F_n(z, \omega) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty; z]}(\xi_i(\omega)), \quad z \in \mathbb{R} \tag{2}$$

where I_A denotes the indicator function of the set A. It is well known that the sequence of empirical distribution functions converges almost surely to the distribution function F under rather general conditions.

2 Problem formulation

For $\mu \in \mathcal{P}(\Xi)$ (representing unknown original distribution) consider a general decision problem

$$\inf_{x \in \mathbb{R}^n} \int_\Xi g(x; \xi) \, \mu(\mathrm{d}\xi) \quad \text{subject to} \quad \mu\big\{\xi \in \Xi : x \in X(\xi)\big\} \geq p_0 \tag{3}$$

where only general conditions on g and X are assumed: $g : \mathbb{R}^n \times \mathbb{R}^s \to \mathbb{R}$, $X : \mathbb{R}^s \rightrightarrows \mathbb{R}^n$ are both measurables, and $p_0 \in [0; 1]$ is a prescribed probability level. Recourse problems and problems with probabilistic constraints are special cases of (3). Denote $\varphi(\mu)$ the optimal value and $\psi(\mu)$ the optimal solution set to the problem (3).

In (3), we consider some estimate $\nu \in \mathcal{P}(\Xi)$ (e. g. an empirical distribution defined by (2)) instead of μ. The next section recalls some Lipschitz and Hölder stability properties of $\varphi(\cdot)$ and $\psi(\cdot)$ with respect to the Wasserstein metrics. It deals particularly with the fixed constraint set $X \subset \mathbb{R}^n$, i. e. with the recourse model of the form

$$\int_\Xi g(x;\xi)\mu(\mathrm{d}\xi) \quad \text{subject to} \quad x \in X. \tag{4}$$

The stability results of Section 3 will next be applied when ν denotes an empirical distribution.

3 Stability results

Theorem 1 *Consider* (4) *where $\mu, \nu \in \mathcal{P}_1(\Xi)$, X is compact, g is uniformly continuous on $\mathbb{R}^n \times \mathbb{R}^s$ and Lipschitz continuous in ξ for all $x \in X_0$ with a constant L independent on x. Then*

$$|\varphi(\mu) - \varphi(\nu)| \leq L\, W_1(\mu, \nu).$$

If in addition X is a convex set and $g(\cdot, \xi)$ is strongly convex function on X with parameter $\sigma > 0$, then

$$\|\psi(\mu) - \psi(\nu)\|^2 \leq \frac{8}{\sigma} L\, W_1(\mu, \nu).$$

PROOF See [3]; for a definition of strongly convex function see [11].

The strong convexity condition of g allows us to consider $\psi(\nu)$ as a (unique) point of $\mathbb{R}^n$ (however it is in fact a singleton).

Theorem 1 is a basic tool to estimate upper bound of difference in optimal value and optimal solution. First, upper bounds of Theorem 1 depend on the model structure by means of the Lipschitz constant L. Especially, the underlying structure can be complex so far as we can guarantee the Lipschitz continuity of g in its random component.

The other component of the upper bound given by Theorem 1 measures the distance between μ and ν by the Wasserstein metrics. We fix our attention to this point now. If the function $g(x;\xi)$ had a separable structure in random element, i. e. $g(x;\xi) = \sum g_i(x;\xi_i)$ where every ξ_i is one-dimensional random variable, then we could directly apply Theorem 1 on each of functions g_i. Here we take advantage of the fact that Wasserstein metrics is easily computable for one-dimensional distributions (see (1)). Consequently, upper bounds for changes in optimal value and optimal solution can be obtained.

4 Empirical distribution

Let μ_n denote an empirical measure with distribution function F_n defined by (2). In [14], limiting distribution for $\sqrt{n}\, W_1(\mu, \mu_n)$ is calculated, where μ is the uniform distribution on

$(0, 1)$. Using the inverse transformation theorem, one could induce a limiting distribution of this statistics for arbitrary distribution; however, our aim is different. The inverse transformation theorem is not always applicable. The problem is that inverse distribution function F^{-1} could not always be given in its explicit form, and other approximation methods take place.

In the latter section, we give a short illustration on how some distributions deals with the Wasserstein metrics applied on themselves and theirs empirical versions. The goal is not to give an explicit formula for limiting distribution (this requires a more sophisticated theory groundwork, see again the book [14]), but to show how large errors one can look for when using different empirical distributions in practical optimization problems.

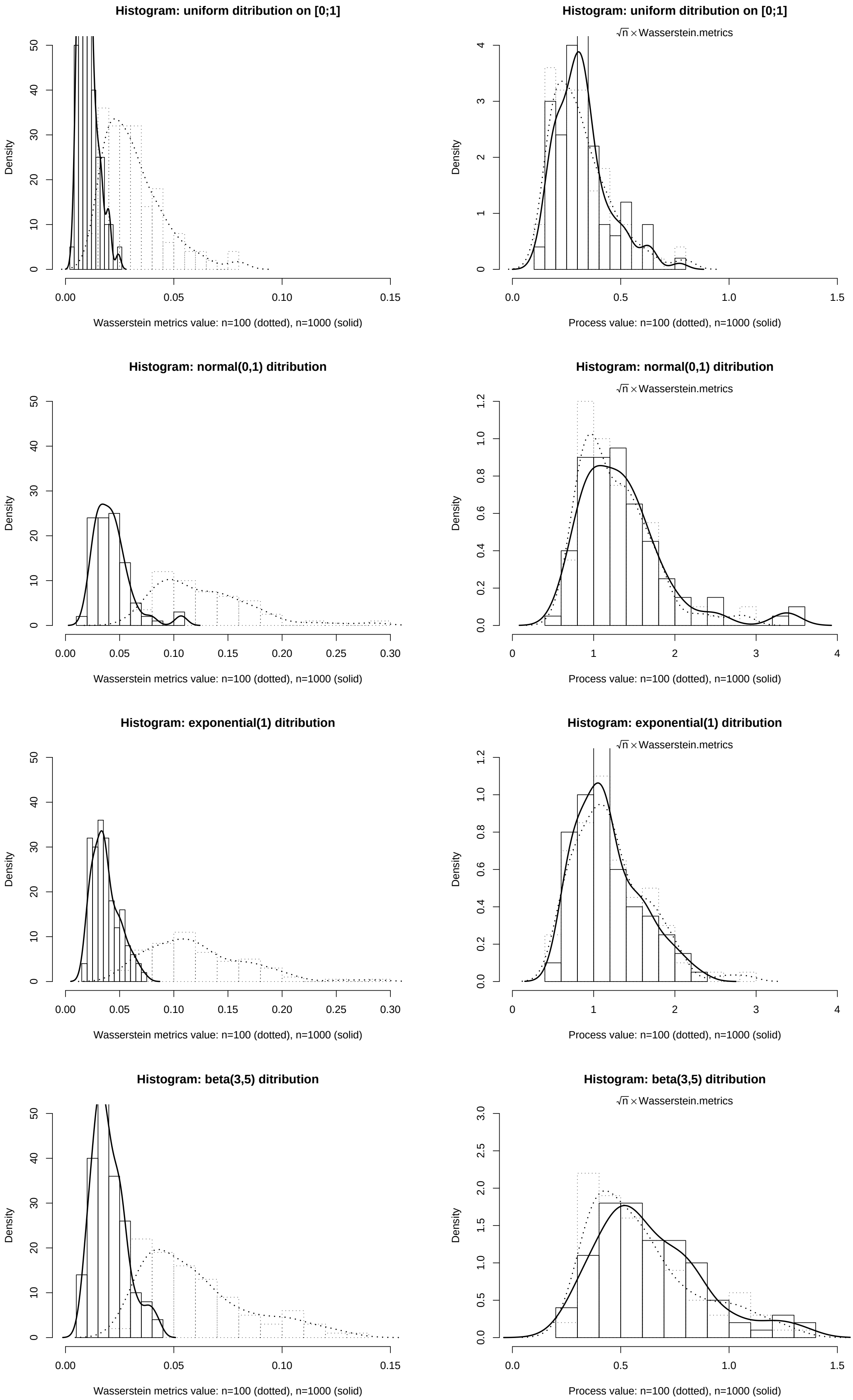
5 Simulations

We use the R programming language for calculations needed to this section. This free statistical tool is very flexible in implementing user-defined procedures. Nevertheless, we have chosen to use built-in but powerfull routines for integration and random number generation:

- (one-dimensional) integral estimate is based on QUADPACK numerical routines, see [8];
- pseudo-random number generator is of the type Mersenne-Twister: a twisted generalized feedback shift register generator, see [7]. Normal distribution is estimated using the inversion procedure and Wichura's algorithm AS 241 for quantile function, see [16].

The exact procedure for estimating $W_1(\mu, \mu_n)$ reads as follows: for a given distribution and a length n, a random sample is generated. Then the empirical distribution is set up and its absolute difference to the original distribution function is integrated giving an estimate for the Wasserstein distance. This procedure is repeated 100 times for each pair (distribution, n) in order to get basic statistical properties of estimates.

The first set of following histograms illustrates a fact that the Wasserstein distance between a distribution and its empirical estimate converges to zero as follows from well-known theoretical results about empirical distributions. The second set of histograms provides distributions of $\sqrt{n}\, W_1(\mu, \mu_n)$ (i. e. rate of the convergence). If one knows the Lipschitz constant of Theorem 1 (derived from the structure of the original model (4)), one can directly apply the theorem to obtain an estimate to the error arising when the original unknown distribution is replaced by its empirical estimate.


Histogram: uniform ditribution on [0;1]
Density
Wasserstein metrics value: n=100 (dotted), n=1000 (solid)
Histogram: uniform ditribution on [0;1]
√n × Wasserstein.metrics
Density
Process value: n=100 (dotted), n=1000 (solid)
Histogram: normal(0,1) ditribution
Density
Wasserstein metrics value: n=100 (dotted), n=1000 (solid)
Histogram: normal(0,1) ditribution
√n × Wasserstein.metrics
Density
Process value: n=100 (dotted), n=1000 (solid)
Histogram: exponential(1) ditribution
Density
Wasserstein metrics value: n=100 (dotted), n=1000 (solid)
Histogram: exponential(1) ditribution
√n × Wasserstein.metrics
Density
Process value: n=100 (dotted), n=1000 (solid)
Histogram: beta(3,5) ditribution
Density
Wasserstein metrics value: n=100 (dotted), n=1000 (solid)
Histogram: beta(3,5) ditribution
√n × Wasserstein.metrics
Density
Process value: n=100 (dotted), n=1000 (solid)

References

[1] Dupačová, J., Gröve-Kuska N., and Römisch, W. (2003), Scenario reduction in stochastic programming: An approach using probability metrics. *Mathematical Programming* Ser. A 95, 493–511.

[2] Dupačová, J., and Römisch, W. (1998), Quantitative stability for scenario-based stochastic programs. In: *Proceedings of the Prague Stochastics '98* (M. Hušková, P. Lachout, and J. Á. Víšek, eds.), Union of Czech Mathematicians and Physicists, Prague, 119–124.

[3] Houda, M. (2002), Probability metrics and the stability of stochastic programs with recourse. *Bulletin of the Czech Econometric Society* 9 (17), 65–77.

[4] Kaňková, V. (1994), On stability in two-stage stochastic nonlinear programming. In: *Asymptotic statistics* (P. Mandl and M. Hušková, eds.), Physica-Verlag, Heidelberg, 329–340.

[5] Kaňková, V. (1997), On the stability on stochastic programming: the case of individual probability constraints. *Kybernetika* 33, 525–546.

[6] Kaňková, V., and Houda, M. (2002), A note on quantitative stability and empirical estimates in stochastic programming. In: *Operations Research Proceedings 2002* (U. Leopold-Wildburger, F. Rendl, and G. Wäsher, eds.), Springer-Verlag, Heidelberg, 413–418.

[7] Matsumoto, M., and Nishimura, T. (1998), A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation* 8, 3–30.

[8] Piessens, R., deDoncker-Kapenga, E., Uberhuber, C., and Kahaner, D. (1983), *QUADPACK: A Subroutine Package for Automatic Integration*. Springer, Verlag.

[9] Rachev, S. T. (1991), *Probability Metrics and the Stability of Stochastic Models*. Wiley, Chichester.

[10] Rachev, S. T., and Römisch, W. (2000), Quantitative stability in stochastic programming: the method of probability metrics. *Mathematics of Operations Research* 27 (4), 792–818.

[11] Rockafellar, R. T., and Wets, R. J.-B. (1997), *Variational Analysis*. Springer, Berlin.

[12] Römisch, W., and Schultz, R. (1991), Distribution sensitivity in stochastic programming. *Mathematical Programming* 50, 197–226.

[13] Römisch, W., and Schultz, R. (1991), Stability analysis for stochastic programs. *Annals of Operation Research* 30, 241–266.

[14] Shorack, G. R., and Wellner, J. A. (1986), *Empirical Processes with Applications to Statistics*. John Wiley&Sons, New York.

[15] Vallander, S. S. (1973), Calculation of the Wasserstein distance between probability distributions on the line. *Theory of Probability and its Applications* 18, 784–786.

[16] Wichura, M. J. (1988), Algorithm AS 241: The Percentage Points of the Normal Distribution. *Applied Statistics* 37, 477–484.

Michal Houda (*houda@karlin.mff.cuni.cz*)
Charles University of Prague, Faculty of Mathematics and Physics
Department of Probability and Mathematical Statistics
Sokolovská 83, 186 75 Praha 8-Karlín, Czech Republic