The generation of scenario trees for multistage stochastic optimization

Georg Pflug

July 2013

Multistage stochastic optimization problems

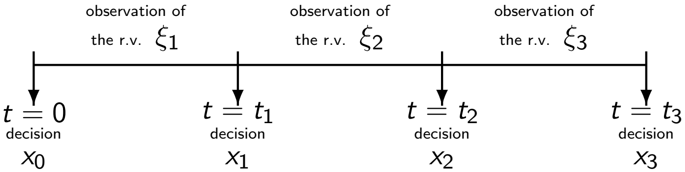
Many real decision problems under uncertainty involve several decision stages:

- hydropower storage and generation management
- thermal electricity generation
- portfolio management
- logistics
- asset/liabilty management in insurance

At each time $t = 0, 1, \ldots, T-1$ a decision x_t can/must be made. We call the sequence $x = (x_0, x_1, \ldots, x_{T-1})$ a *strategy*. The costs of the strategy x is expressed in terms of a cost function, which depends also on some random parameters (the scenario process) $\xi = (\xi_1, \ldots, \xi_T)$ defined on some probability space $(\Omega, \mathcal{F}, P)$

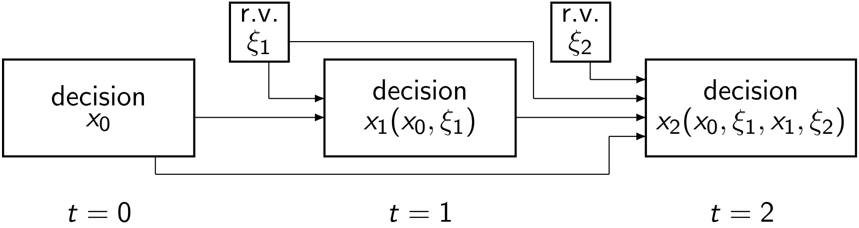
$$Q(x_0, \xi_1, x_1, \ldots, x_{T-1}, \xi_T).$$

Multistage decisions



Decisions can only be made on the basis of the available information. For this reason, we assume that a filtration $\mathfrak{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_T = \mathcal{F})$ is defined in $(\Omega, \mathcal{F}, P)$ such that $\xi_t \lhd \mathcal{F}_t$ (ξ_t is measurable w.r.t. $\mathcal{F}_t$).

Multistage stochastic decision processes



Decisions are functions of past observations and past decisions.

The Decision Problem

The final objective is to minimize a functional $\mathcal{R}$ of the stochastic cost function, such as the expectation, a quantile or some other functional $\mathcal{R}$

$$(Opt)\left\|\begin{array}{l}\text{Minimize in } x_0, x_1(\xi_1), \ldots, x_{T-1}(\xi_1, \ldots, \xi_{T-1}) : \\ \mathcal{R}[Q(x_0, \xi_1, \ldots, x_{T-1}, \xi_T)] \\ \ \text{s.t. } x \lhd \mathfrak{F} \\ \ \text{and possibly other constraints on } x_0, \ldots, x_{T-1} : x \in \mathbb{X}\end{array}\right.$$

$x \lhd \mathfrak{F}$ means that $x_t \lhd \mathcal{F}_t$, i.e. that the decisions are

nonanticipative.

Scenario Processes

Scenario process are often multidimensional:

- hydropower storage and generation management: rainfall, electricity spotprices, demand
- thermal electricity generation: fuel prices, spotprices, demand
- portfolio management: asset prices
- logistics: demands at the nodes of the logistic network
- asset/liabilty management in life insurance: Asset prices, mortality pattern, demand for contracts

Sometimes the available information is more than just the cost relevant process ξ, e.g. rainfall in other areas which allow better estimates for the rainfall in the hydrostorage area. That is why we distinguish between the filtration $\mathfrak{F}$ and the information generated by ξ:

$$\sigma(\xi) \subseteq \mathfrak{F}.$$

Approximations

In order to numerically solve the multiperiod stochastic optimization problem, the stochastic process (ξ_t) must be approximated by a simple stochastic process $\tilde{\xi}_t$, which takes only a small number of values. Likewise the filtration $\mathfrak{F}$ must be approximated by a smaller one $\tilde{\mathfrak{F}}$ such that $\sigma(\tilde{\xi}) \subseteq \tilde{\mathfrak{F}}$.

$$\tilde{F}(\tilde{x}_1, \ldots, \tilde{x}_{T-1}) = \mathcal{R}[Q(\tilde{x}_0, \tilde{\xi}_1, \tilde{x}_1, \ldots, \tilde{x}_{T-1}, \tilde{\xi}_T)]$$

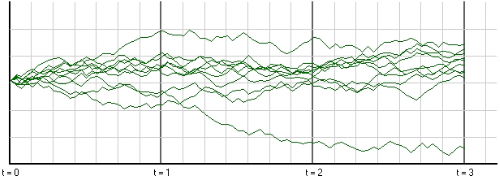
$$(\widetilde{Opt}) \left\| \begin{array}{l} \text{Minimize in } \tilde{x}_0, x_1(\tilde{\xi}_1), \ldots, \tilde{x}_{T-1}(\tilde{\xi}_1, \ldots, \tilde{\xi}_{T-1}) : \\ \mathcal{R}[Q(\tilde{x}_0, \tilde{\xi}_1, \ldots, \tilde{x}_{T-1}, \tilde{\xi}_T)] \\ \ \text{s.t. } \tilde{x} \lhd \tilde{\mathfrak{F}} \\ \ \text{and possibly other constraints} \tilde{x} \in \tilde{\mathbb{X}}. \end{array} \right.$$

Approximation of stochastic decision processes

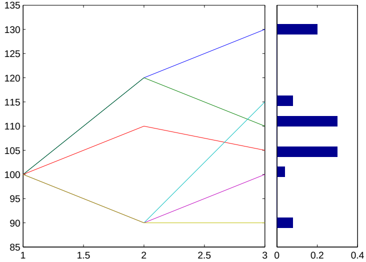
$$
\begin{tikzcd}[column sep=huge, row sep=huge]
\begin{array}{c}\text{original problem}\\ (Opt)\end{array} \arrow[r, dashed, "\text{solution}"] \arrow[dd, "\text{approximation}"'] & \begin{array}{c}x^*,\\ \text{the solution of } (Opt)\end{array} \arrow[d, leftrightarrow, "\text{approximation error } e = F(x^*) - F(x^+)"] \\
 & x^+ = \pi_{\mathbb{X}}(\tilde{x}^*) \\
\begin{array}{c}\text{approximate problem}\\ (\widetilde{Opt})\end{array} \arrow[r, "\text{solution}"'] & \begin{array}{c}\tilde{x}^*,\\ \text{the solution of } (\widetilde{Opt})\end{array} \arrow[u, "\text{extension}"']
\end{tikzcd}
$$

The scenario generation problem

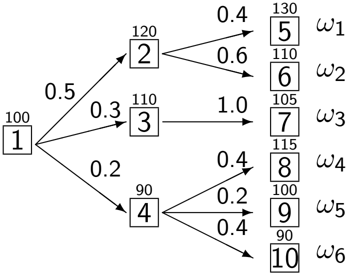
Out of a scenario process



we want to make a scenario tree



Node-valuated (scenarios) and arc-valuated (probabilities) trees



$P(\omega_1) = 0.2$

$P(\omega_2) = 0.3$

$P(\omega_3) = 0.3$

$P(\omega_4) = 0.08$

$P(\omega_5) = 0.04$

$P(\omega_6) = 0.08$

$\mathcal{N}_0$ $\quad$ $\mathcal{N}_1$ $\quad$ $\mathcal{N}_2 = \Omega$, the sample space

An exemplary finite tree process $\nu = (\nu_0, \nu_1, \nu_2)$ with nodes $\mathcal{N} = \{1, \ldots 10\}$ and leaves $\mathcal{N}_2 = \{5, \ldots 10\}$ at $T = 2$ stages. The filtrations, generated by the respective atoms, are

$\mathcal{F}_2 = \sigma\left(\{\omega_1\}, \{\omega_2\}, \ldots \{\omega_6\}\right)$,

$\mathcal{F}_1 = \sigma\left(\{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5, \omega_6\}\right)$ and

$\mathcal{F}_0 = \sigma\left(\{\omega_1, \omega_2, \ldots \omega_6\}\right)$

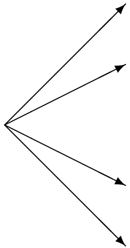
- Scenario generation is not just a heuristic method, but a part of approximation theory.
- The fundament of the mathematics of scenario generation is the notion of distances between probability measures (i.e. multivariate distributions and stochastic processes).
- The theory of probability quantization deals with the approximation of probability distributions by those sitting on finitely points.
- The approximation error can be bounded by the distance between the scenario models.
- Statistical results on the quality of approximation are available

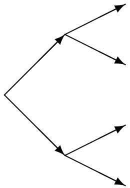
The approximation dilemma

The approximation should be coarse enough to allow an efficient numerical solution but also fine enough to make the approximation error small. It is therefore of fundamental interest to understand the relation between the complexity and the approximation quality of approximative models.

We quantify the approximation error by a new distance concept, the *nested distance* for scenario processes and the information structure.

Single-,two- and multistage

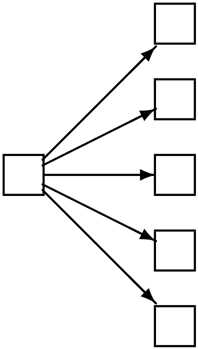


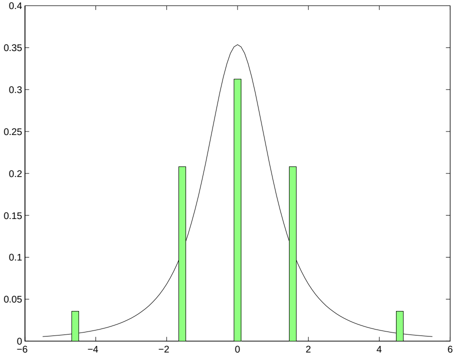


Single- or twostage Multistage

Single- or twostage scenario generation just involves to generate a list of values and probabilities. No information-related aspect arises. For multistage problems, the tree structure, which encodes the information structure is very important.

The one-period case





P original probability measure, $\tilde{P}$ discrete approximation

$$\tilde{P} : \begin{array}{l|cccc} \text{probabilities} & p_1 & p_2 & \cdots & p_S \\ \hline \text{values} & z_1 & z_2 & \cdots & z_S \end{array}$$

Distances of Probability measures

Distances for probability measures are typically defined as

$$\mathrm{d}_{\mathcal{H}}(P,\tilde{P}) = \sup\{|\int h(w)\, dP(w) - \int h(w)\, d\tilde{P}(w)| : h \in \mathcal{H}\},$$

where $\mathcal{H}$ is a class of functions.

- The *uniform distance* (Kolmogorov-Smirnov distance)

$$\mathrm{d}_U(P,\tilde{P}) \;=\; \sup\{|P(-\infty,a] - \tilde{P}(-\infty,a]| : a \in \mathbb{R}^d\}$$

- The *Kantorovich distance*

$$\mathrm{d}_1(P,\tilde{P}) = \sup\{\int h\, dP - \int h d\tilde{P} : |h(u) - h(v)| \le \|u - v\|\}.$$

- The *Fortet-Mourier* distance

$$\mathrm{d}_{FM_p}(P,\tilde{P}) = \sup\{\int h\, dP - \int h d\tilde{P} : L_p(h) \le 1,$$

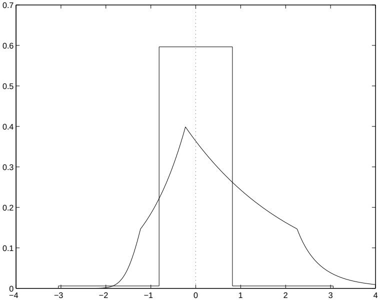
where

$$L_p(f) = \inf\{L : |h(u) - h(v)| \le L|u - v| \max(1, |u|^{p-1}, |v|^{p-1})\}.$$

The *moment matching semidistance* is not a distance:

$$d_{MM}(P, \tilde{P}) = \sup\{ \int w^p \, dP(w) - \int w^p \, d\tilde{P}(w) : 1 \leq p \leq M \}$$

Closedness of moments does not tell anything about the closedness of the corresponding distributions.



Two densities g_1 and g_2 with identical first four moments.

Integral inequalities: The uniform distance

Hlawka-Koksma Inequality:

$$|\int h(u)\, dP(u) - \int h(u)\, d\tilde{P}(u)| \leq \mathrm{d}_U(P, \tilde{P}) \cdot V(h).$$

where $V(h)$ is the Hardy-Krause variation of h: Let $V^{(M)}(h) = \sup \sum_{J_1,\dots,J_n \text{ is a partition by rectangles } J_i} |\Delta_{J_i}(h)|$, where $\Delta_J(h)$ is the sum of values of h at the vertices of J, where adjacent vertices get opposing signs. The Hardy-Krause Variation of h is $\sum_{m=1}^{M} V^{(m)}(h)$.

In the univariate situation, if K is a monotonic function, then

$$\mathrm{d}_U(K(\xi), K(\tilde{\xi})) = \mathrm{d}_U(\xi, \tilde{\xi}).$$

Here, the distance between random variables is defined as the distance between their distributions.

Using the quantile transform, the univariate approximation problem reduces to approximate the uniform [0,1] distribution.

Integral inequalities: The Kantorovich distance

Let $L(h)$ be the Lipschitz constant of the function h:

$$L(h) = \sup\{\frac{|h(u) - h(v)|}{d(u,v)} : u \neq v\}.$$

$$|\int h\, dP - \int h d\tilde{P}| \leq L(h) \cdot \mathrm{d}_1(P, \tilde{P}).$$

Theorem (Kantorovich-Rubinstein). Dual version of Kantorovich-distance:

$$\begin{aligned} \mathrm{d}_1(P,\tilde{P}) &= \inf\{\mathbb{E}(\mathrm{d}(X,Y) : (X,Y) \text{ is a bivariate r.v. with} \\ &\qquad \text{given marginal distributions } P \text{ and } \tilde{P}\}. \end{aligned}$$

Generalization: The Wasserstein-distance of order r

$$\begin{aligned} \mathrm{d}_r(P,\tilde{P}) &= \inf\{\left(\int \mathrm{d}(u,v)^r\, d\pi(u,v)\right)^{1/r} : \pi \text{ is a probability distributi} \\ &\qquad \text{on } \Xi \times \tilde{\Xi} \text{ with given marginal distributions } P \text{ and } \tilde{P}\}. \end{aligned}$$

Remark. If both measures sit on a finite number of mass points $\{z_1, z_2, \ldots, z_s\}$, then $\mathrm{d}_r^r(P, \tilde{P})$ is the optimal value of the following linear optimization problem:

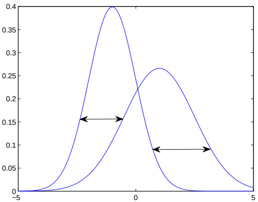
$$\left\| \begin{array}{ll} \text{Minimize } \sum_{i,j} pi_{ij} d_{ij} & \\ \sum_i \pi_{ij} = \tilde{P}_j & \text{for all } j \\ \sum_j \pi_{ij} = P_i & \text{for all } i \end{array} \right.$$

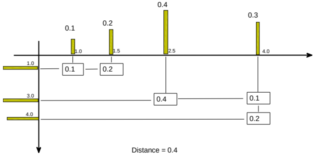
For $r = 1$ this problem has a dual

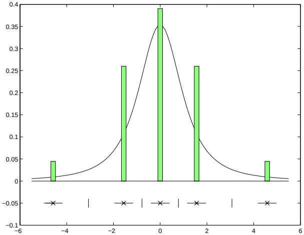
$$\left\| \begin{array}{ll} \text{Maximize } \sum_i y_i (P_i - \tilde{P}_i) & \\ y_i - y_j \leq d_{ij} & \text{for all } i, j \end{array} \right.$$

Here P_i resp. $\tilde{P}_i$ is the mass sitting on z_i and $d_{ij} = \mathrm{d}(z_i, z_j)$.

Interpretation as mass transportation/facility location problem







Kantorovich distance: Average distance to the next facility
Uniform distance: Worst case distance to the next facility (works only on bounded spaces)

Historic remarks

The distance d_1 was introduced by Kantorovich in 1942 as a distance in general spaces. In 1948, he established the relation of this distance (in $\mathbb{R}^m$) to the mass transportation problem formulated by Gaspard Monge in 1781 (*Monge's mass transportation problem*). In 1969, L. N. Wasserstein –unaware of the work of Kantorovich – this distance for using it for convergence results of Markov processes and one year later R. L. Dobrushin used and generalized this distance and initiated the name Wasserstein distance. S. S. Vallander studied the special case of measures in $\mathbb{R}^1$ in 1974 and this paper made the name Wasserstein metric popular. Modern books have been written by Rachev and Rüschendorf (1998) and Villani (2003).

Implications of closedness in Wasserstein distance

Assume that $X \sim P$ and $\tilde{X} \sim \tilde{P}$. Then

1. $\left|\mathbb{E}|X|^p - \mathbb{E}|\tilde{X}|^p\right| \le$
   $p \cdot \mathrm{d}_r\left(P, \tilde{P}\right) \cdot \max\left\{\mathbb{E}^{\frac{r-1}{r}}\left[|X|^{r \cdot \frac{p-1}{r-1}}\right], \mathbb{E}^{\frac{r-1}{r}}\left[|\tilde{X}|^{r \cdot \frac{p-1}{r-1}}\right]\right\}$,
2. $|\mathbb{E}(X^p) - \mathbb{E}(X^p)| \le$
   $p \cdot \mathrm{d}_r\left(P, \tilde{P}\right) \cdot \max\left\{\mathbb{E}^{\frac{r-1}{r}}\left[|X|^{r \cdot \frac{p-1}{r-1}}\right], \mathbb{E}^{\frac{r-1}{r}}\left[|\tilde{X}|^{r \cdot \frac{p-1}{r-1}}\right]\right\}$ for p an integer,
3. $\left|\mathbb{E}X^2 - \mathbb{E}\tilde{X}^2\right| \le 2 \cdot \mathrm{d}_2\left(P, \tilde{P}\right) \cdot \max\left\{\mathbb{E}^{\frac{1}{2}}\left[X^2\right], \mathbb{E}^{\frac{1}{2}}\left[\tilde{X}^2\right]\right\}$,
4. $\left|\mathbb{E}|X|^r - \mathbb{E}|\tilde{X}|^r\right| \le$
   $r \cdot \mathrm{d}_r\left(P, \tilde{P}\right) \cdot \max\left\{\mathbb{E}^{\frac{r-1}{r}}\left[|X|^r\right], \mathbb{E}^{\frac{r-1}{r}}\left[|\tilde{X}|^r\right]\right\}$ and
5. $\left|\mathbb{E}|X|^p - \mathbb{E}|\tilde{X}|^p\right| \le$
   $p \cdot \mathrm{d}_2\left(P, \tilde{P}\right) \cdot \max\left\{\mathbb{E}^{\frac{1}{2}}\left[|X|^{2(p-1)}\right], \mathbb{E}^{\frac{1}{2}}\left[|\tilde{X}|^{2(p-1)}\right]\right\}$,

where $p \ge 1$ and $r > 1$.

Alternative metrics on $\mathbb{R}$

It is not necessary to measure the distance in $\mathbb{R}$ by $d(u,v) = |u - v|$. Alternatively one may use the distance

$$d_\chi(u,v) = |\chi(u) - \chi(v)|$$

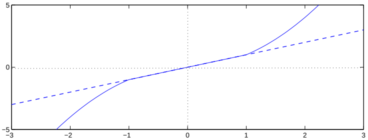
where χ is a strictly monotone function on $\mathbb{R}$. A distance on $\mathbb{R}^m$ ca be defined as

$$d(u,v) = \sum_{i=1}^{m} |\chi_i(u_i) - \chi_i(v_i)|.$$

A typical example for χ is

$$\chi_q(u) = \begin{cases} u & |u| \leq 1 \\ sign(u) \cdot |u|^q & |u| > 1 \end{cases}$$

Relation to the Fortet-Mourier metric



The following relation holds for $q \geq 1$:

$$\frac{1}{q} d_1(P_1, P_2 | d_{\chi_q}) \leq d_{FM}(P_1, P_2) \leq 2 d_1(P_1, P_2 | d_{\chi_q})$$

The approximation w.r.t. the Fortet-Mourier distance can be traced back to the approximation w.r.t. the Kantorovich distance through the transformation

Let G be a distribution function on $\mathbb{R}$:

- Choose q: Transform G with $\chi^q : G^{1/q} = G \circ \chi^{1/q}$
- Approximate $G^{1/q}$ by $\tilde{G}^{1/q}$ by minimizing the Kantorovich distance
- Backtransformation: $\tilde{G} = \tilde{G}^{1/q} \circ \chi^q$

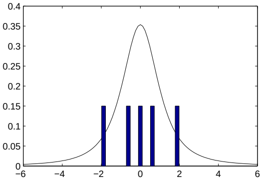
A comparison

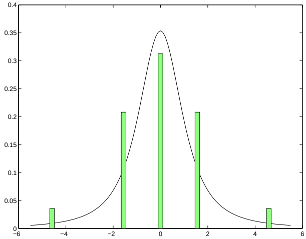
Suppose we want to approximate the t-distribution with 2 degrees of freedom by a probability measure sitting on five points. Using the uniform distance one gets the solution $\tilde{P}_1$

probability	0.2	0.2	0.2	0.2	0.2
value	-1.8856	-0.6172	0	0.6172	1.8856

Using the Kantorovich distance one gets $\tilde{P}_2$

probability	0.0446	0.2601	0.3906	0.2601	0.0446
value	-4.58	-1.56	0	1.56	4.58





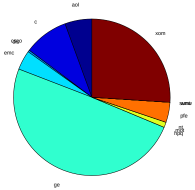
Case study: Comparison of scenario generation methods

- 13 (somewhat randomly selected) assets: AOL, C, CSCO, DIS, EMC, GE, HPQ, MOT, NT, PFE, SUNW, WMT, XOM
- Weekly data from January 1993 to January 2003.
- Rolling horizon: asset allocation, backtracking
- Optimize MAD and AVaR with full data and approximations

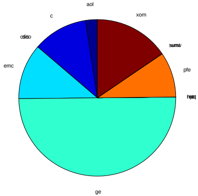
(R. Hochreiter and G. Pflug)

Empirical results: MAD

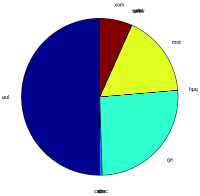
Example (Asset Allocation): Data 01/1993-01/1995, $n = 50$



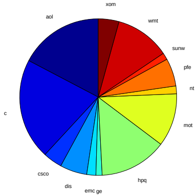
Full dataset



Kantorovich



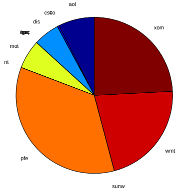
Moment matching



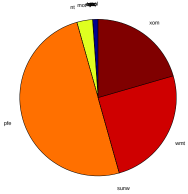
Sobol QMC

Empirical results: AVaR$_\alpha$

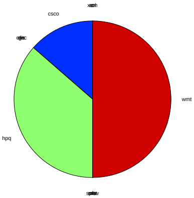
Example (Asset Allocation): Data 01/1999-01/2001, $\alpha = 0.1, n = 50$



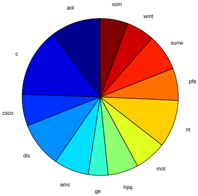
Full dataset



Kantorovich



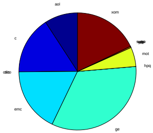
Moment matching



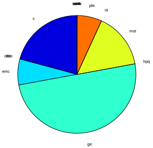
Sobol QMC

Empirical results: AVaR_α

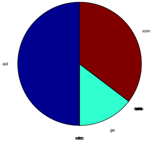
Example (Asset Allocation): Data 01/1993-01/1995,
$\alpha = 0.1, S = 50$



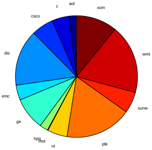
Full dataset



Kantorovich



Moment matching



Sobol QMC

Optimal discretizations

The basic problem: Let d be some distance for probability measures and let $\mathcal{P}_s$ be the family of probability distributions sitting on at most s points. For $P \in \mathcal{P}_s$, one wants to find the *quantization error*

$$q_{s,d}(P) = \inf\{d(P,Q) : Q \in \mathcal{P}_s\} \tag{1}$$

and the *optimal quantization* set (could consists of several probability distributions)

$$\mathcal{Q}_{s,d}(P) = \text{argmin } \{d(P,Q) : Q \in \mathcal{P}_s\} \tag{2}$$

(if it exists).

An example for exact optimality

Let P be a Laplace distribution with density $g(u) = \frac{1}{2}\exp(-|u|)$.

$$q_{s,d_1}(P) = \begin{cases} \log(1+\frac{2}{s}) & s \text{ even} \\ \frac{2}{s+1} & s \text{ odd} \end{cases}$$

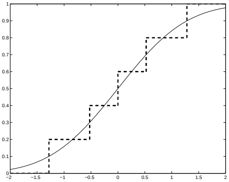
Here d_1 is the Kantorovich distance belonging to the Euclidean norm. The optimal distributions of points is also known.

Optimality w.r.t. uniform distance

The optimal approximation of a continuous probability p with distribution function G by a distribution sitting on at most s mass points $z_1, \ldots, z_s$ with probabilities $p_1, \ldots, p_s$ w.r.t. the uniform distance is given by

$$z_i = G^{-1}(\frac{2i-1}{2s}), \quad p_i = 1/s, i = 1, \ldots, s.$$

The distance is $1/s$.



A similar result for multivariate distributions (copulas on $[0,1]^m$) is unknown.

Asymptotically optimal discretizaitons

Theorem. Suppose that P has a density g such that $\int |u|^{r+\delta} g(u)\, du < \infty$ for some $\delta > 0$. Then

$$\inf_s s^{1/m} q_{s,d_p}(P) = \bar{q}^{(m)}_{d_p} \left[\int_{\mathbb{R}^m} [g(x)]^{m/(m+1)}\, dx \right]^{(m+1)/m}.$$

where $q^{(m)}_{d_p,s} = \inf_s s^{1/M} q_{s,d_p}(\mathcal{U}[0,1]^s))$ (exact value unknown).

References: Book by Graf and Luschgy, work of Gilles Pages, Klaus Poetzelberger and many others.

Monte Carlo sampled scenarios

Let $X_1, X_2, \ldots, X_s$ be an i.i.d. sequence distributed according to P. Then the Monte Carlo approximation is

$$\hat{P}_s = \frac{1}{s} \sum_{i=1}^{s} \delta_{X_i}.$$

The MC approximation in uniform distance.

Theorem(Kolmogorov). An asymptotic result: Let P be the uniform distribution in [0,1] and $X_1, X_2, \ldots$ be an i.i.d. sequence from a P. Then

$$\lim_{s\to\infty} P\{\sqrt{s} d_U(P, \hat{P}_s) > t\} = 2 \sum_{k=1}^{\infty} (-1)^{k+1} \exp(-2k^2 t^2).$$

Theorem(Dvoretzky, Kiefer, Wolfowitz inequality). A nonasymptotic result:

$$\mathbb{P}\{d_U(P, \hat{P}) \geq \epsilon/\sqrt{S}\} \leq 58 \exp(-2\epsilon^2)$$

The MC approximation in Kantorovich distance.

Theorem (Graf and Luschgy). Let $X_1, X_2, \ldots$ be an i.i.d. sequence from a distribution with density g in $\mathbb{R}^m$. Then

$$\lim_{s \to \infty} P\{s^{1/m} d_1(P, \hat{P}_s) > t\} = \int (1 - \exp(-t^m b_m g(x)) g(x)\, dx.$$

where $b_m = \frac{2\pi^{m/2}}{m\Gamma(m/2)}$.

Large deviations

Theorem (Boley, Guilin and Villani). Suppose that there is an $\alpha > 0$ such that $\int \exp(\alpha d^2(x,y)))\,P(dx) < \infty$. Then there is a $\lambda > 0$ and a $N_0 > 0$ such that for all $\lambda' > \lambda$, $m' > m$ and $n \geq N_0 \max(\epsilon^{-m'-2}, 1)$

$$P\{d_1(\hat{P}_s, P) \geq \epsilon\} \leq \exp(-\frac{\lambda'}{2} n\epsilon^2).$$

Theorem (Boley, Guilin, Villani). Let $d(x,y) = \|x-y\|$. Suppose that $\int \exp(\alpha\|x\|)\,P(dx) < \infty$. Then for $m' < m$, there exist constants k and N_0 such that for $\epsilon > 0$ and $n \geq N_0 \max(\epsilon^{-(2r+m')}, 1)$

$$P\{d_r(\hat{P}_s, P) \geq \epsilon\} \leq \exp(-Kn^{1/r} \min(\epsilon, \epsilon^2)).$$

Distances for stochastic processes (nested distributions)

If (Ξ_1, d_1) and (Ξ_2, d_2) are Polish spaces then so is the Cartesian product $(\Xi_1 \times \Xi_2)$ with metric

$$d^2((u_1, u_2), (v_1, v_2)) = d_1(u_1, v_1) + d_2(u_2, v_2).$$

Consider some metric d on $\mathbb{R}^m$, which makes it Polish (it needs not to be the Euclidean one). Then we define the following spaces

$$\begin{aligned}
\Xi_1 &= (\mathbb{R}^m, d) \\
\Xi_2 &= (\mathbb{R}^m \times \mathcal{P}_1(\Xi_1, d), d^2) = (\mathbb{R}^m \times \mathcal{P}_1(\mathbb{R}^m, d), d^2) \\
\Xi_3 &= (\mathbb{R}^m \times \mathcal{P}_1(\Xi_2, d), d^2) = (\mathbb{R}^m \times \mathcal{P}_1(\mathbb{R}^m \times \mathcal{P}_1(\mathbb{R}^m, d), d^2), d^2) \\
&\vdots \\
\Xi_T &= (\mathbb{R}^m \times \mathcal{P}_1(\Xi_{T-1}, d), d^2)
\end{aligned}$$

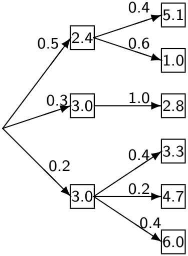
All spaces $\Xi_1, \ldots, \Xi_T$ are Polish spaces and they may carry probability distributions.

Definition. A probability distribution $\mathbb{P}$ with finite first moment on Ξ_T is called a *nested distribution of depth* T.
For any nested distribution $\mathbb{P}$, there is an embedded multivariate distribution P, which has lost the information structure. The projection from the nested distribution to the embedded distribution is not injective!

Notation for discrete distributions:

$$\begin{array}{l} \text{probabilities:} \\ \text{values:} \end{array} \quad \left[\begin{array}{ccc} 0.3 & 0.4 & 0.3 \\ \hline 3.0 & 1.0 & 5.0 \end{array} \right]$$

Examples for nested distributions



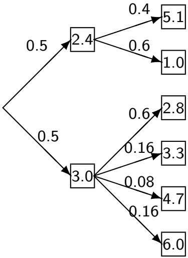
$$\mathbb{P} = \left[\begin{array}{ccc} 0.2 & 0.3 & 0.5 \\ \hline \\ 3.0 & 3.0 & 2.4 \\ \left[\begin{array}{ccc} 0.4 & 0.2 & 0.4 \\ \hline 6.0 & 4.7 & 3.3 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ \hline 2.8 \end{array} \right] & \left[\begin{array}{cc} 0.6 & 0.4 \\ \hline 1.0 & 5.1 \end{array} \right] \end{array} \right]$$

The embedded multivariate, but non-nested distribution of the scenario process can be gotten from it:

$$\left[\begin{array}{cccccc} 0.08 & 0.04 & 0.08 & 0.3 & 0.3 & 0.2 \\ \hline \\ 3.0 & 3.0 & 3.0 & 3.0 & 2.4 & 2.4 \\ 6.0 & 4.7 & 3.3 & 2.8 & 1.0 & 5.1 \end{array} \right]$$

Evidently, the embedded multivariate distribution has lost the information about the nested structure. If one considers the filtration generated by the scenario process itself and forms the pertaining nested distribution, one gets

$$\left[\begin{array}{cc} 0.5 & 0.5 \\ \hline 3.0 & 2.4 \\ \left[\begin{array}{cccc} 0.16 & 0.08 & 0.16 & 0.6 \\ \hline 6.0 & 4.7 & 3.3 & 2.8 \end{array}\right] & \left[\begin{array}{cc} 0.6 & 0.4 \\ \hline 1.0 & 5.1 \end{array}\right] \end{array}\right]$$



Distances between nested distributions

Since a nested distribution is a distribution on the metric space Ξ_T (which consists of values and distributions) the notion of Kantorovich distance makes sense. If $\mathbb{P}$ and $\tilde{\mathbb{P}}$ are two nested distributions on Ξ_T, then the distance $\mathsf{dl}(\tilde{\mathbb{P}}, \mathbb{P})$ is well defined. This distance makes sense, even if one process is discrete and the other is not.

Theorem. Let $\mathbb{P}, \tilde{\mathbb{P}}$ be nested distributions and $P, \tilde{P}$ be the pertaining multiperiod distributions. Then

$$d(P, \tilde{P}) \le \mathsf{dl}(\mathbb{P}, \tilde{\mathbb{P}}).$$

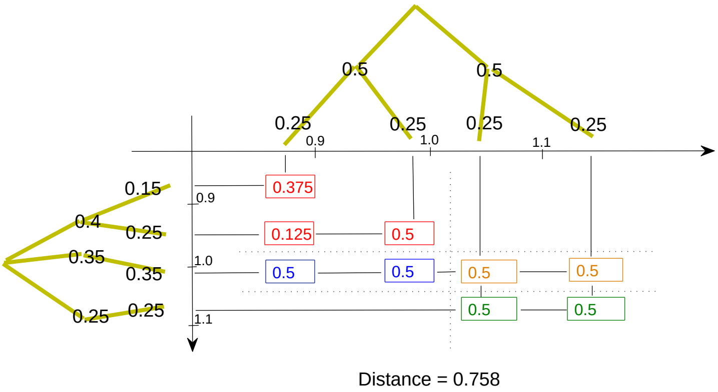
Alternative characterization of the nested distance

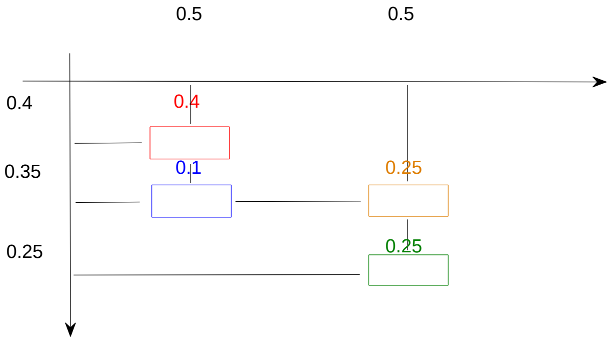
Theorem. For two nested distributions $\mathbb{P} := (\Xi, \mathcal{F}, P)$, $\tilde{\mathbb{P}} := \left(\tilde{\Xi}, \tilde{\mathcal{F}}, \tilde{P}\right)$ and a distance function on $d\colon \Xi \times \Xi' \to \mathbb{R}$ the *nested distance of order* $r \geq 1$ – denoted $\mathrm{dl}_r\left(\mathbb{P}, \tilde{\mathbb{P}}\right)$ – is the optimal value of the optimization problem

$$
\begin{array}{lll}
\underset{(\text{in } \pi)}{\text{minimize}} & \left(\int d\left(\xi, \tilde{\xi}\right)^r \pi\left(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}\right)\right)^{\frac{1}{r}} & \\
\text{subject to} & \pi\left(M \times \tilde{\Xi} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right) = P\left(M \mid \mathcal{F}_t\right) & (M \in \mathcal{F}_T) \\
& \pi\left(\Xi \times N \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right) = \tilde{P}\left(N \mid \tilde{\mathcal{F}}_t\right) & (N \in \tilde{\mathcal{F}}_T)
\end{array}
$$

where the minimum is among all bivariate probability measures $\pi \in \mathcal{P}\left(\Xi \times \Xi'\right)$, which are measures on the product sigma algebra $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$. We will refer to the nested distance also as *process distance*, or *multistage distance*. The nested distance dl_2 (order $r = 2$), with d a weighted Euclidean distance is referred to as *quadratic nested distance*.

The nested distance: illustration




0.5
0.5
0.4
0.4
0.35
0.1
0.25
0.25
0.25

How to calculate the nested distance

The Wasserstein distance between discrete trees can be calculated by solving the a linear program

$$\begin{array}{rl} \text{minimize} & \\ \text{(in } \pi) & \sum_{i,j} \pi_{i,j} \cdot d^r_{i,j} \\ \text{subject to} & \sum_{j \succ n} \pi\left(i, j \middle| m, n\right) = P\left(i \middle| m\right) \quad (m \prec i,\ n), \\ & \sum_{i \succ m} \pi\left(i, j \middle| m, n\right) = \tilde{P}\left(j \middle| n\right) \quad (n \prec j,\ m), \\ & \pi_{i,j} \geq 0 \text{ and } \sum_{i,j} \pi_{i,j} = 1, \end{array}$$

where again $\pi_{i,j}$ is a matrix defined on the leave nodes ($i \in \mathcal{N}_T$, $j \in \mathcal{N}'_T$) and $m \in \mathcal{N}_t$, $n \in \mathcal{N}'_t$ are arbitrary nodes. The conditional probabilities $\pi\left(i, j \middle| m, n\right)$ are given by

$$\pi\left(i, j \middle| m, n\right) = \frac{\pi_{i,j}}{\sum_{i' \succ m, j' \succ n} \pi_{i',j'}}.$$

Example for the nested distance between a continuous process and a tree

Let

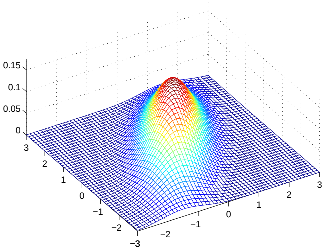
$$\mathbb{P} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right).$$

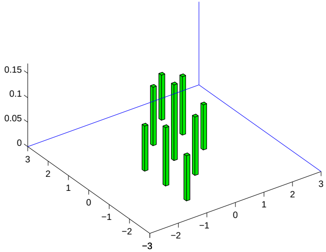
and

$$\tilde{\mathbb{P}} =$$

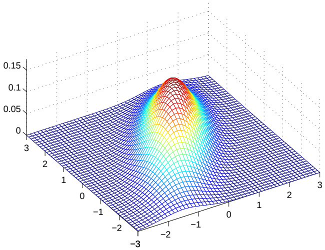
$$\left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline -1.029 & 0.0 & 1.029 \\ \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline -2.058 & -1.029 & 0.0 \end{array}\right] & \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline -1.029 & 0.0 & 1.029 \end{array}\right] & \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline 0.0 & 1.029 & 2.058 \end{array}\right] \end{array}\right]$$

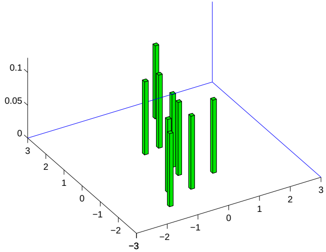
The nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 0.82$.
The distance of the multiperiod distributions is $d(P, \tilde{P}) = 0.68$.





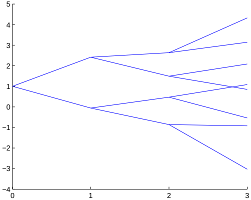
The nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 0.82$.
The distance of the multiperiod distributions is $d(P, \tilde{P}) = 0.68$.



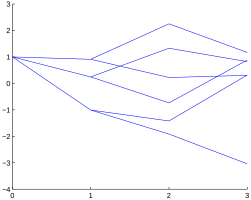


The nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 1.12$.
The distance of the multiperiod distributions is $d(P, \tilde{P}) = 0.67$.

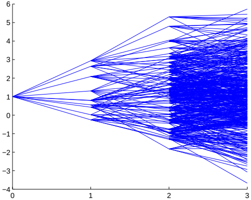
Examples of nested distances



$\mathbb{P}^{(1)}$: tree 1



$\mathbb{P}^{(2)}$: tree 2



$\mathbb{P}^{(3)}$: tree 3

$$\mathrm{dl}(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) = 3.90; \qquad d(P^{(1)}, P^{(2)}) = 3.48$$

$$\mathrm{dl}(\mathbb{P}^{(1)}, \mathbb{P}^{(3)}) = 2.52; \qquad d(P^{(1)}, P^{(3)}) = 1.77$$

$$\mathrm{dl}(\mathbb{P}^{(2)}, \mathbb{P}^{(3)}) = 3.79; \qquad d(P^{(2)}, P^{(3)}) = 3.44$$

The main approximation result

Let $\mathcal{Q}_L$ be the family of all real valued cost functions $Q(x_0, y_1, x_1, \dots, x_{T-1}, y_T)$, defined on $\mathbb{X}_0 \times \mathbb{R}^{n_1} \times \mathbb{X}_1 \times \dots \times \mathbb{X}_{T-1} \times \mathbb{R}^{n_T}$ such that

- $x = (x_0, \dots, x_{T-1}) \mapsto Q(x_0, y_1, x_1, \dots, x_{T-1}, y_T)$ is convex for fixed $y = (y_1, \dots, y_T)$ and
- $y_t \mapsto Q(x_0, y_1, x_1, \dots, x_{t_1}, y_T)$ is Lipschitz with Lipschitz constant L for fixed x.

Consider the optimization problem $(Opt(\mathbb{P}))$

$$v_Q(\mathbb{P}) := \min\{\mathbb{E}_P[Q(x_0, \xi_1, x_1, \dots, x_{T-1}, \xi_T)] : x \lhd \mathfrak{F}, x \in \mathbb{X}\},$$

where $\mathbb{X}$ is a convex set and $\mathbb{P}$ is the nested distribution of the scenario process.

An approximative problem $(Opt(\tilde{\mathbb{P}}))$ is given by

$$v_Q(\tilde{\mathbb{P}}) := \min\{\mathbb{E}_{\tilde{P}}[Q(x_0, \tilde{\xi}_1, x_1, \dots, x_{T-1}, \tilde{\xi}_T)] : x \lhd \tilde{\mathfrak{F}}, x \in \mathbb{X}\},$$

where $\tilde{\mathbb{P}}$ is the nested distribution of the approximative scenario process.

Theorem. For Q in $\mathcal{Q}_L$

$$|v_Q(\mathbb{P}) - v_Q(\tilde{\mathbb{P}})| \leq L \cdot \mathsf{dl}(\mathbb{P}, \tilde{\mathbb{P}}).$$

Remarks.

- The bound is sharp: Let $\mathbb{P}$ and $\tilde{\mathbb{P}}$ be two nested distributions on $[\Xi, \mathsf{dl}]$. Then there exists a cost function $Q(\cdot) \in \mathcal{H}_1$ such that

$$v_Q(\mathbb{P}) - v_Q(\tilde{\mathbb{P}}) = \mathsf{dl}(\mathbb{P}, \tilde{\mathbb{P}}).$$

- The inequality

$$|v_Q(\mathbb{P}) - v_Q(\tilde{\mathbb{P}})| \leq L \cdot d(\mathbb{P}, \tilde{\mathbb{P}}),$$

where d is the multivariate Kantorovich distance, does NOT hold.

Distortion functionals

Let G_Y be the distribution function of Y. Then the distortion functional $\mathcal{R}_\sigma$ with distortion density σ is defined as

$$\mathcal{R}_\sigma(Y) = \int_0^1 \sigma(u) G_Y^{-1}(u)\, du$$

A special example is the average value-at-risk, which has distortion density

$$\sigma_\alpha(u) = \begin{cases} 0 & u < \alpha \\ \frac{1}{1-\alpha} & u \geq \alpha \end{cases}$$

An extension of the main result

Theorem. Let $\mathcal{R}_\sigma$ be a distortion risk functional with bounded distortion, $\sigma \in L^\infty$.
Consider the optimization problem $(Opt(\mathbb{P}))$

$$v_{Q,\mathcal{R}_\sigma}(\mathbb{P}) := \min\{\mathcal{R}_{\sigma,\mathbb{P}}[Q(x_0,\xi_1,x_1,\dots,x_{T-1},\xi_T)] : x \lhd \mathfrak{F}, x \in \mathbb{X}\},$$

where $\mathbb{X}$ is a convex set and $\mathbb{P}$ is the nested distribution of the scenario process.
An approximative problem $(Opt(\tilde{\mathbb{P}}))$ is given by

$$v_{Q,\mathcal{R}}(\tilde{\mathbb{P}}) := \min\{\mathcal{R}_{\sigma,\tilde{\mathbb{P}}}[Q(x_0,\tilde{\xi}_1,x_1,\dots,x_{T-1},\tilde{\xi}_T)] : x \lhd \tilde{\mathfrak{F}}, x \in \mathbb{X}\},$$

where $\tilde{\mathbb{P}}$ is the nested distribution of the approximative scenario process.
Then

$$|v_{Q,\mathcal{R}_\sigma}(\mathbb{P}) - v_{Q,\mathcal{R}_\sigma}(\tilde{\mathbb{P}})| \leq L \cdot \|\sigma\|_\infty \cdot \mathsf{dl}_1\left(\mathbb{P},\tilde{\mathbb{P}}\right).$$

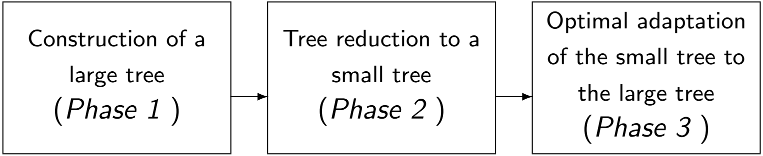
Scenario tree generation

- Dupacova, Consiglio, Wallace (2000). Clustering and sequential sampling, importance sampling
- Dupacova, Groewe-Kuska, Roemisch (2003). Scenario generation using probability metrics
- Heitsch, Roemisch (2009). Scenario tree reduction
- Heitsch, Roemisch (2011). Filtration distance

Scenario tree generation using the nested distance

Suppose that $\xi_1, \ldots, \xi_T$ is a random scenario process and that a random number generator is available which generates the conditional distributions $\xi_{t+1}|\xi_1, \ldots, \xi_t$.
The tree generation algorithm has three phases

- In phase 1 a large tree is generated using a stochastic gradient method for optimal discretization of the conditional distributions.
- In phase 2, the large tree is reduced to an acceptable size.
- In phase 3, the smaller tree is brought as close as possible to the original large tree.

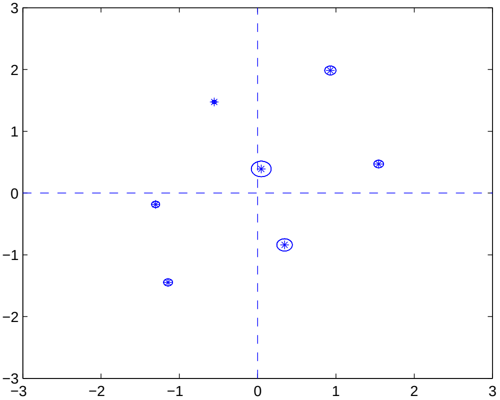

Construction of a large tree (Phase 1)
Tree reduction to a small tree (Phase 2)
Optimal adaptation of the small tree to the large tree (Phase 3)

Phase 1: Facility location by stochastic gradient search

Suppose that we can generate an i.i.d. sequence of random values $\xi^{(k)}$. The *stochastic approximation* algorithm is

1. Initialize
$$\begin{aligned}\tilde{\Xi}^{(0)} &= \{\tilde{\xi}_i^{(0)} : 1 \le i \le s\} \\ \tilde{p}_i^{(0)} &= 1/s \qquad \text{for } 1 \le i \le s\end{aligned}$$
2. Observe the next random value $\xi^{(k)}$
3. Find $j \in \{1, 2, \ldots, s\}$ such that $\xi^{(k)}$ is closest to $\tilde{\xi}_j^{(k)}$.
4. Set $\tilde{\xi}_j^{(k+1)} = \frac{k}{k+1}\tilde{\xi}_j^{(k)} + \frac{1}{k+1}\xi_j^{(k)}$ and leave all other points unchanged.
5. Estimate
$$\tilde{p}_j^{(k+1)} = \frac{k\tilde{p}_j^{(k)} + 1}{k+1} \qquad \tilde{p}_i^{(k+1)} = \frac{k\tilde{p}_i^{(k)}}{k+1} \qquad \text{for } i \neq j$$
6. Set $k := k + 1$ and goto 2.

Example



The best 7 points to represent a twodimensional normal distribution.

Incorporating constraints

Sometimes it is needed to incorporate constraints such as conditions for the expectation to avoid arbitrage in investment models.
Writing the previous algorithm as

$$P^{(k+1)} = \frac{k}{k+1} P^{(k)} + \frac{1}{k+1} (\delta_{\xi_j^{(k+1)}} - \delta_{\xi_j^{(k)}})$$

this algorithm can be modified to

$$P^{(k+1)} = \mathrm{proj}_{\mathcal{P}} \left[\frac{k}{k+1} P^{(k)} + \frac{1}{k+1} (\delta_{\xi_j^{(k+1)}} - \delta_{\xi_j^{(k)}}) \right].$$

Phase 2: Scenario tree reduction by merging subtrees

- **Step 1 – Choice of the subtrees to be merged.** Let a tree $\mathbb{P}$ be given. At each level t the nested distance between all subtrees is calculated. Let $\mathbb{P}_1$and $\mathbb{P}_2$ be the two subtrees at stage t which are closest to each other and should be merged into one. To do so, we use the algorithm Merging trees.
- **Step 2 – Merging trees.**
  1. For merging two trees into one, the new value ξ_1 at the new root is the mean of the two values of the two old roots.
  2. For the successors of the two roots the averaging algorithm with parameter p is used. Suppose that the selected pairs of nodes are
     $$(i_1, j_1), \dots (i_m, j_m).$$
     Then, in a recursive step, the subtrees with roots i_1 and j_1 have to be merged, as well as all other pairs i_2 and j_2 up to i_m and j_m.
- **Stop or continue**. If the new tree is small enough, stop. Otherwise choose another level t and another pair of close subtrees to be merged into one by going to Step 1.

Phase 3: The tree adaptation algorithm (Kovacevic and Pichler)

- **Step 1– Initialization**
  Set $k \leftarrow 0$, and let ξ^0 be process quantizers with related transport probabilities $\pi^0(i,j)$ between scenario i of the original $\mathbb{P}$-tree and scenario $\tilde{\xi}^0_j$ of the approximating $\mathbb{P}'$-tree; $\mathbb{P}^0 := \tilde{\mathbb{P}}$.
- **Step 2 – Improve the quantizers**
  Find improved quantizers $\tilde{\xi}^{k+1}_j$:
  - In case of the quadratic Wasserstein distance (Euclidean distance and Wasserstein of order $r = 2$) set

$$\tilde{\xi}^{k+1}(n_t) := \sum_{m_t \in \mathcal{N}_t} \frac{\pi^k(m_t, n_t)}{\sum_{m_t \in \mathcal{N}_t} \pi^k(m_t, n_t)} \cdot \xi_t(m_t),$$

  - or find the barycenters by applying the steepest descent method, or the limited memory BFGS method.

- **Step 3 – Improve the probabilities**
  Find the new transportation plan π^* using the new quantizers $\tilde{\xi}$ and calculate all conditional probabilities $\pi^{k+1}(\cdot,\cdot|m,n) = \pi^*(\cdot,\cdot|m,n)$, the unconditional transport probabilities $\pi^{k+1}(\cdot,\cdot)$ and the distance $\mathrm{dl}_r^{k+1} = \mathrm{dl}_r\left(\mathbb{P},\tilde{\mathbb{P}}\right)$.
- **Step 4**
  Set $k \leftarrow k+1$ and continue with Step 2 if

  $$\mathrm{dl}_r^{k+1} < \mathrm{dl}_r^{k} - \varepsilon,$$

  where $\varepsilon > 0$ is the desired improvement in each cycle k.
  Otherwise, set $\tilde{\xi}^* \leftarrow \tilde{\xi}^k$, define the measure

  $$\tilde{P}^{k+1} := \sum_j \delta_{\tilde{\xi}_j^{k+1}} \cdot \sum_i \pi^{k+1}(i,j),$$

  for which $\mathrm{dl}_r\left(\mathbb{P},\mathbb{P}^{k+1}\right) = \mathrm{dl}_r^{k+1}$ and stop.

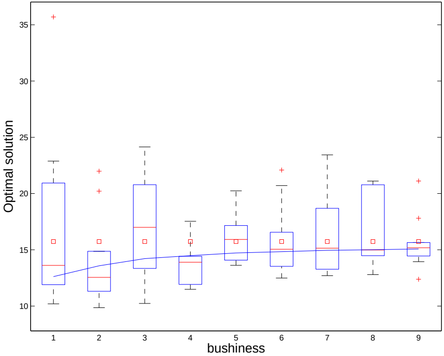
In case of the quadratic nested distance ($r=2$) and the Euclidean distance the choice $\varepsilon = 0$ is possible.

Computational experience

Stages	4	5	5	6	7	7
Nodes of the initial tree	53	309	188	1,365	1,093	2,426
Nodes of the approx. tree	15	15	31	63	127	127
Time/ sec.	1	10	4	160	157	1,044

Monte Carlo sampling versus optimal quantization using nested distances

Monte Carlo sampling: results vary and the box-plots are shown
Optimal quantization: blue line; true value: red boxes



An inventory control problem (the multistage newsboy problem)

Approximation at work

Reducing the nested distance by making the tree bushier.