Decision making under uncertainty: data-driven modeling

Georg Ch. Pflug

October 26, 2016

Decision making under uncertainty

It is well known that for many decision problems under uncertainty, the multistage formulation of a decision problem outperforms a repeated myopic single- or two-stage approach. Examples of such problems include

- Portfolio optimization
- Asset-liability management
- Pension fund management
- Managing energy portfolios (production and trading)
- Hydrostorage management (turbining and pumping)
- Transportation and logistics
- Supply Chains and inventory control

Scenario processes

For appropriate modeling, we need time series data, such as

- Portfolio optimization: asset prices, option prices
- Asset-liability managment: liability processes
- Pension fund management: Retirement process and mortality of customers
- Energy portfolios: spot prices, future prices, fuel prices, demand patterns
- Hydrostorage management: inflows, spot prices, pumping prices, demands
- Transportation and logistics: transportation costs, demands
- Supply Chains and inventory Control: inventory holding costs, order costs, demands

The problem formulation

We consider a multistage stochastic optimization problem of the form

$$Opt(\mathbb{P}): \quad v^*(\mathbb{P}) = \min\{\mathcal{R}_{\mathbb{P}}[Q(x,\xi)] \,:\, x \lhd \mathfrak{F}; \mathbb{P} \sim (\Omega, \mathfrak{F}, P, \xi)\} \tag{1}$$

where Ω is a probability space, $\mathfrak{F}$ is a filtration on Ω, i.e. an increasing sequence of sigma-algebras

$$\mathfrak{F} = (\mathcal{F}_0 = (\Omega, \emptyset), \mathcal{F}_1, \dots, \mathcal{F}_T = 2^{\Omega}), \tag{2}$$

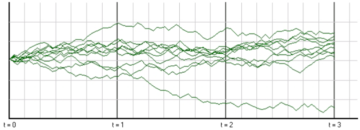
$\xi = (\xi_0, \dots, \xi_T)$ is a stochastic *scenario process* adapted to the filtration $\mathfrak{F}$ (in symbol $\xi \lhd \mathfrak{F}$). The notation $Q(x,\xi)$ is a short form of

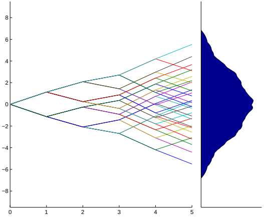
$$Q(x,\xi) = Q(x_0, \xi_1, x_1, \xi_2, \dots, \xi_T, x_T).$$

$\mathcal{R}$ is a risk functional.

Discretizations

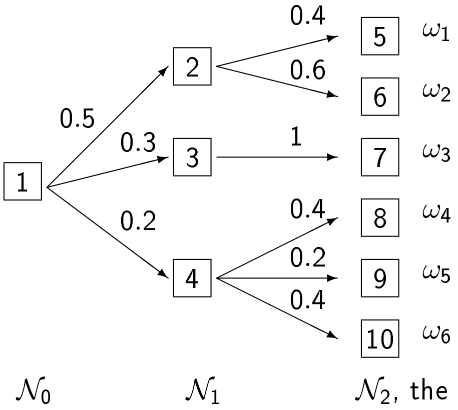
While in reality many scenario processes evolve in continuous time and continuous space, the numerical treatment requires to discretize time and space. The appropriate discretized object is a *scenario tree*.





We solve the multistage decision problem on a tree.

Trees represent discrete stochastic processes



$P(\omega_1) = 0.2$

$P(\omega_2) = 0.3$

$P(\omega_3) = 0.3$

$P(\omega_4) = 0.08$

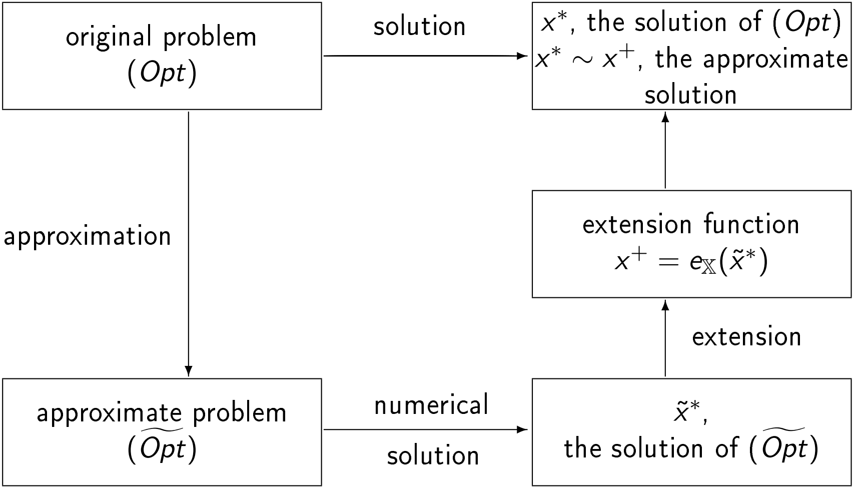
$P(\omega_5) = 0.04$

$P(\omega_6) = 0.08$

$\mathcal{N}_2$, the sample space

An exemplary finite tree process 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 $\mathcal{F}_0 = \sigma\left(\{\omega_1, \omega_2, \ldots \omega_6\}\right)$

Approximations



original problem (Opt)

solution

x^*, the solution of (Opt) $x^* \sim x^+$, the approximate solution

approximation

extension function $x^+ = e_{\mathbb{X}}(\tilde{x}^*)$

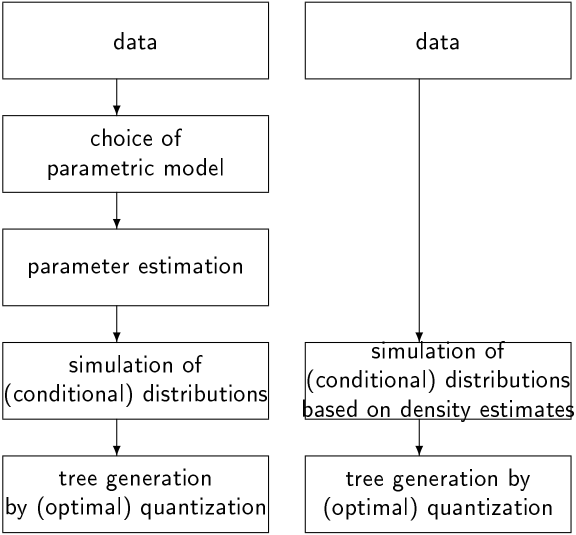
extension

approximate problem $(\widetilde{Opt})$

numerical solution

$\tilde{x}^*$, the solution of $(\widetilde{Opt})$

From data to tree models



data

choice of parametric model

parameter estimation

simulation of (conditional) distributions

tree generation by (optimal) quantization

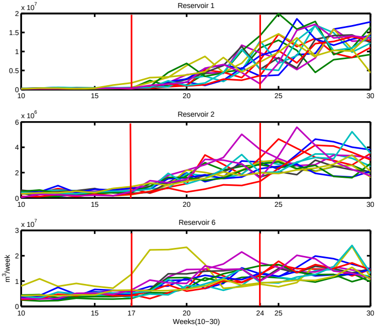
data

simulation of (conditional) distributions based on density estimates

tree generation by (optimal) quantization

The parametric (left) and the nonparametric approach (right)

Illustration: Inflow data



Alternatives:

- parametric way: Identify a $SARMA(1,2),(2,2)_{52}$ model with normal errors and estimate all parameters
- nonparametric way: Find the optimal discretizations based on conditional density estimates

The Wasserstein distance (Monge's transportation distance)

Let (Ξ, d) be a metric space, typically $\mathbb{R}^m$, let $\mathcal{P}_1(\Xi, d)$ be the family probability measures P on (Ξ, d) such that

$$\int d(u, u_0)\, dP(z) < \infty$$

for some $u_0 \in \Xi$.
$\mathcal{P}_1$ is a complete separable metric space under the Wasserstein/Kantorovich transportation distance d_1:

$$\begin{aligned} \mathrm{d}_1(P_1, P_2) \quad = \quad & \sup\{ | \int f(u)\, dP_1(u) - \int f(u)\, dP_2(u) | : \\ & |f(u) - f(v)| \leq d(u, v) \} \end{aligned}$$

Here the supremum is over all Lipschitz(1) functions w.r.t the basic distance d.

The Kantorovich-Rubinstein Theorem

Theorem.

$$\begin{aligned}
\mathrm{d}_1(P_1,P_2) &= \inf\{\mathbb{E}(d(X,Y) : (X,Y) \text{ is a bivariate r.v. with} \\
&\quad \text{given marginal distributions } P_1 \text{ and } P_2\}.
\end{aligned}$$

A generalization of the Kantorovich distance is the Wasserstein distance of order r

$$\begin{aligned}
\mathrm{d}_r^r(P_1,P_2) &= \inf\{\mathbb{E}(d(X,Y)^r : (X,Y) \text{ is a bivariate r.v. with} \\
&\quad \text{given marginal distributions } P_1 \text{ and } P_2\}.
\end{aligned}$$

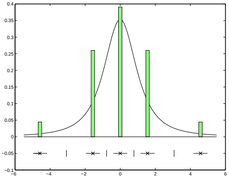
$$d_r^r(P_1,P_2) = \inf_\pi \iint_{\Xi\times\tilde{\Xi}} \mathrm{d}(x,y)^r \pi(\mathrm{d}x,\mathrm{d}y),$$

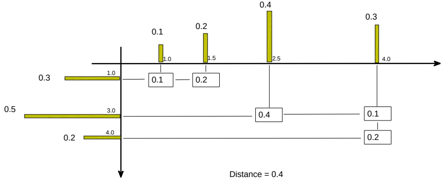
where π is a probability measure with given marginals P_1 and P_2,

$$\begin{aligned}
\pi(A\times\Xi) &= P_1(A) \text{ and} \\
\pi(\Xi\times B) &= P_2(B).
\end{aligned}$$

The infimum is attained. The minimizer π is called the *optimal transportation plan*.

Interpretation as facility location/mass transportation problem





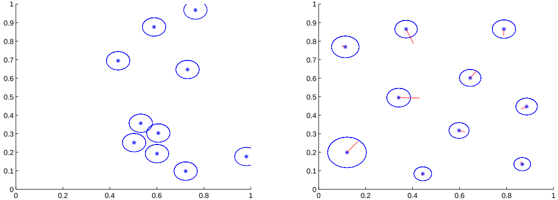
If both probabilities are discrete, then the calculation of the Wasserstein distance is a linear program.

Optimal quantization

Ideally one would like to solve

$$\min\{\mathrm{d}_r(P, \tilde{P}_s). \tilde{P}_s \text{ sits on at most } s \text{ points}\}$$

This optimal facility location problem is a NP-hard problem, but can anyway be solved by stochastic (quasi-) gradient methods (at least to local optimality). An often used alternative, is to take just take a random sample from the probability distribution P, but ...



Left: A MC sample form the Uniform$[0,1]^2$ distribution
Right: A nearly optimal discretization form the same distribution

The approximation quality of Monte Carlo sampling

Let $\hat{P}_s$ be the empirical distribution based on a sample of size s from the underlying distribution P. Then, for any square integrable function f.

$$|\int f(u)\, d\hat{P}_s(u) - \int f(u)\, dP(u)| = O_P(s^{-1/2})$$

irrespective of the dimension m of the sample space $\Xi \subseteq \mathbb{R}^m$. This formula is however not uniform in f. To the contrary: If uniformity in f is required, strong conditions on the uniformity set F have to be imposed.

Theorem. (Talagrand,1994)

$$P\{\sup_{f \in F} |\int f(u)\, d\hat{P}_s(u) - \int f(u)\, dP(u)| \geq M \cdot s^{-1/2}\} \leq \frac{C}{M} \left(\frac{M^2}{V} \right)^v e^{-2M^2}$$

where F is a family of functions which can be covered by at most $(V/\epsilon)^v$ balls of radius ϵ.

The family Lip(1) is not of finite covering type.

Large deviations

Theorem. (Graf and Luschgy, 2000) Let P have density g in $\mathbb{R}^m$. Then

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

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

Theorem. (Boley, Guilin and Villani, 2007). Let P be a measure on $\mathbb{R}^m$ endowed with metric d. Suppose that there is an $\alpha > 0$ such that $\int \exp(\alpha d^2(u,0))) \, P(du) < \infty$. Then there is a $\lambda > 0$ and a $s_0 > 0$ such that for all $m' > m$ and $s \geq s_0 \max(\epsilon^{-m'-2}, 1)$

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

The approximation quality of optimal discretizations

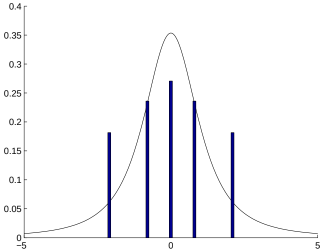
Theorem. (Zador, 1982) Let $\tilde{P}_s$ be the optimal discretization with s points of P, Then

$$\lim_{s\to\infty} s^{1/m} \mathrm{d}_1(P, \tilde{P}_s) = \|g\|_{m/(m+1)} \inf s^{1/m} \mathrm{d}_1(U[0,1]^m, \tilde{U}_s).$$

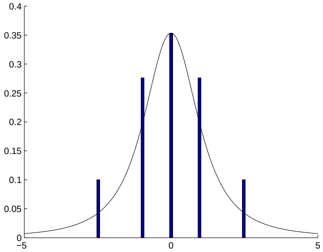
Bounds are known for the latter constant.

The flexibility of the Wasserstein distance

The basic distance d on (Ξ, d) determines the set of Lipschitz(1) functions and therefore the optimal discretization.
Example: Nearly optimal discretization of a $t(2)$ distribution:



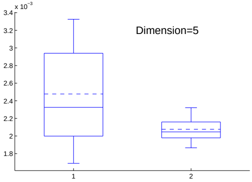
$$d(u, v) = |u - v|$$



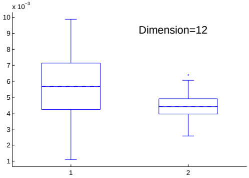
$$d(u, v) = |u^5 - v^5|$$

Monte Carlo versus (nearly) optimal quantization

A Lipschitz function $f(u_1, \ldots, u_m)$ was considered in $[0,1]^m$ and its integral was calculated using Monte Carlo and optimal discretization. Not only the function f was considered, but also permuted versions $f_\sigma = f(u_{\sigma(1)}, \ldots, u_{\sigma(m)})$ for 120 permutations σ. Ideally, all estimated integrals should be the same. The Monte Carlo estimates are not uniform w.r.t. the functions f_σ, while the optimal discretization leads to uniform approximations. Left: MC values for the functions f_σ, Right: integrations using the nearly optimal discretizations for all f_σ.



Dimension m=5



Dimension m=12

Multistage generalization: the nested distance

Definition. Let (Ξ, d) be a metric space (typically $\Xi \subseteq \mathbb{R}^m$) and let

$$\mathbb{P} := \left(\Xi, (\Sigma_t)_{t=0,\dots T}, P\right) \text{ and } \tilde{\mathbb{P}} := \left(\Xi, (\tilde{\Sigma}_t)_{t=0,\dots T}, \tilde{P}\right)$$

be filtered probability spaces.

Definition. (G.P. 2009, G.P., A. Pichler 2014) The *nested distance* of order $r \geq 1$ is

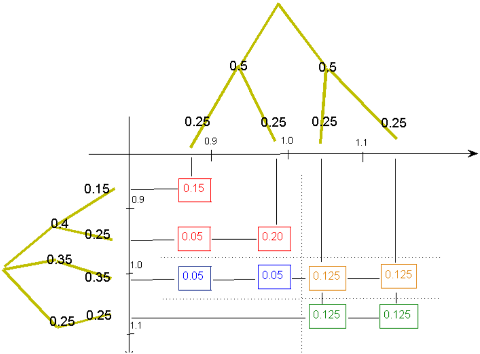
$$\mathsf{dl}_r(\mathbb{P}, \tilde{\mathbb{P}})^r = \inf_{\pi} \iint_{\Xi \times \tilde{\Xi}} \mathsf{d}(x, y)^r \pi(\mathrm{d}x, \mathrm{d}y),$$

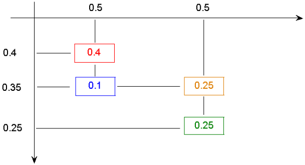
where π is a probability measure with conditional marginals P and $\tilde{P}$, i.e.,

$$\begin{aligned} \pi\left(A \times \Xi \middle| \Sigma_t \otimes \tilde{\Sigma}_t\right) &= P\left(A \middle| \Sigma_t\right) \text{ and} \\ \pi\left(\Xi \times B \middle| \Sigma_t \otimes \tilde{\Sigma}_t\right) &= \tilde{P}\left(B \middle| \tilde{\Sigma}_t\right) \text{ for all } t = 0, \dots T, \end{aligned}$$

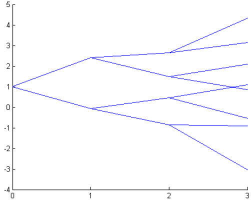
where $A \in \Sigma_T$ and $B \in \tilde{\Sigma}_T$.

Transportation distance between stochastic processes

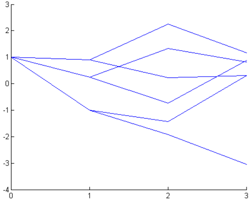




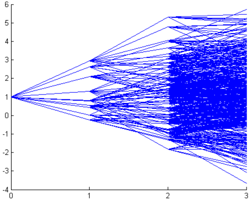
Examples of nested distances



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



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



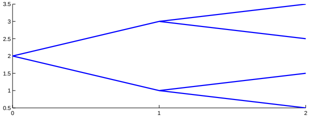
$\mathbb{P}^{(3)}$: tree 3

$$d(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) = 3.90$$

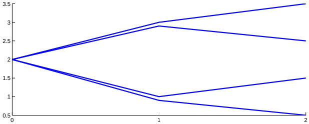
$$d(\mathbb{P}^{(1)}, \mathbb{P}^{(3)}) = 2.52$$

$$d(\mathbb{P}^{(2)}, \mathbb{P}^{(3)}) = 3.79$$

Nested distance $\neq$ Multivariate Wasserstein Distance



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



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

$$\mathsf{d}(\mathbb{P}^{(A)}, \mathbb{P}^{(B)}) = 0.525; d(P^{(A)}, P^{(B)}) = 0.05$$

Stochastic optimization problems are continuous w.r.t. the nested distance

Theorem. (A. Pichler, G.P. 2009)
Let $\mathbb{P} := \big(\Xi, (\Sigma_t)_{t=0,\dots T}, P\big)$ ($\tilde{\mathbb{P}} := \big(\tilde{\Xi}, (\tilde{\Sigma}_t)_{t=0,\dots T}, \tilde{P}\big)$, resp.) be a filtered probability space. Consider the multistage stochastic optimization problem

$$v(\mathbb{P}) := \inf \left\{ \mathbb{E}_P Q(\xi, x) : \, x \lhd \Sigma \right\},$$

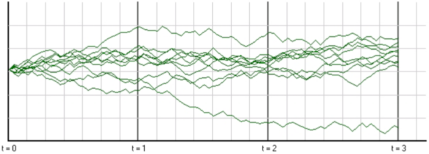
where Q is convex in x for any ξ fixed, and Lipschitz with constant L in ξ for any x fixed. Then

$$\left| v(\mathbb{P}) - v(\tilde{\mathbb{P}}) \right| \leq L \cdot \mathrm{d}_r(\mathbb{P}, \tilde{\mathbb{P}})$$

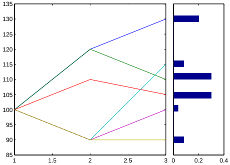
for every $r \geq 1$.
The constraint $x \lhd \Sigma$ is shorthand for $x_t \lhd \Sigma_t$ for all $t = 1, \dots T$, where $x = (x_t)_{t=1}^T$ is the stochastic decision process: x must be adapted to Σ, i.e. *nonanticipative*.

The scenario generation problem

Out of a (discrete time) scenario process $\xi = (\xi_0, \dots, \xi_T)$ with deterministic ξ_0



we want to make a finite scenario tree $\tilde{\xi} = (\tilde{\xi}_0, \dots, \tilde{\xi}_T)$



From data to tree models

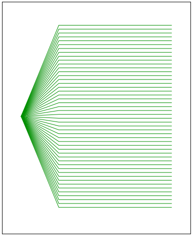
The empirical measure of the i.i.d time series observations

$$\begin{aligned}\xi_1 &= (\xi_{1,1}, \dots \xi_{T,1}) \\ &\dots \\ \xi_n &= (\xi_{1,n}, \dots \xi_{T,n})\end{aligned}$$

is $\hat{P}_n := \frac{1}{n}\sum_{i=1}^n \delta_{\xi_i} = \frac{1}{n}\sum_{i=1}^n \delta_{(\xi_{1,i},\dots\xi_{T,i})}$ on $\mathbb{R}^m$, where each $\xi_i = (\xi_{1,i}, \dots \xi_{T,i})$ is an observation of an entire sample path. The empirical measure is a special case of a discrete measure. Discrete measures are dense w. r. t. the Wasserstein distance in the space of measures satisfying an adequate moment constraint (Bolley, 2008). Under the same conditions, the empirical measure converges a.s. to the true one.

Smoothing is necessary

Notice that the empirical process based on n trajectories can be graphically represented by a "fan", a non-branching tree. The empirical process on $\mathbb{R}^{mT}$ converges (under a moment condition) a.s. in the multivariate Wasserstein distance to the underlying distribution, but *not* in the nested distribution. For the convergence of the conditional distributions some smoothing is necessary.



Conditional density estimation

In order to describe conditional density estimation suppose that X is the main variable (in $\mathbb{R}^{m_x}$) of interest and ξ represents all conditioning variables (in $\mathbb{R}^{m_\xi}$). Then the kernel estimate for the conditional density based on a sample (X_i, ξ_i) is given by

$$\hat{f}_n(x|\xi) = \sum_{i=1}^{n} \frac{\frac{1}{h_\xi^{m_\xi}} k\left(\frac{\xi-\xi_i}{h_\xi}\right)}{\sum_{j=1}^{n} \frac{1}{h_\xi^{m_\xi}} k\left(\frac{\xi-\xi_j}{h_\xi}\right)} \cdot \frac{1}{h_x^{m_x}} k\left(\frac{x-X_i}{h_x}\right)$$

where (h_x, h_ξ) are the respective bandwidths (Parzen-Rosenblatt estimator). This estimator can be rewritten as

$$\hat{f}_n(x|\xi) = \sum_{i=1}^{n} w_i(\xi) \cdot k_{h_x}(x - X_i), \text{ where } w_i(\xi) := \frac{k\left(\frac{\xi-\xi_i}{h_\xi}\right)}{\sum_{j=1}^{n} k\left(\frac{\xi-\xi_j}{h_\xi}\right)}$$

are the weights corresponding to the partial observation X_i.

Density estimation and Wasserstein distance

Lemma. For a translation invariant distance d it holds that

$$\mathrm{d}_r(\tilde{P} * k_h,\ P) \leq \mathrm{d}_r(\tilde{P}, P) + \kappa_r^{1/r} \cdot h,$$

where $\kappa_r = \int \|x\|^r\, k(x)\mathrm{d}x$ is the r^{th}-absolute moment of the kernel k.

Theorem. Let P be a measure on $\mathbb{R}^m$ with density f. Suppose the kernel is Lipschitz with constant $\|k\|_{Lip}$ and supported in the unit ball, $\{k(\cdot) > 0\} \subseteq \{\|\cdot\| \leq 1\}$. Then the kernel density estimator $\hat{f}_n$ corresponding to $\hat{P}_n * k_{h_n}$ satisfies

$$\left\|\hat{f}_n - f\right\|_\infty \leq \delta_f(h) + \frac{\|k\|_{Lip}}{h^{m+1}} \mathrm{d}_r(P, \hat{P}_n)$$

for every $r \geq 1$. Here $\delta_f(h) := \sup_{\{\|x-y\| \leq h\}} |f(x) - f(y)|$ is the modulus of continuity of the density f. (see Bolley et al. 2007).

The main result

Theorem. Introduce the notation $\xi_{0:t} := (\xi_0, \dots \xi_t)$ for a substring of $(\xi_0, \dots \xi_T)$. Suppose that

1. the conditions of the previous two Theorems hold, and
2. the measure P is conditionally Lipschitz, i.e.,
   $d\left(P(\cdot|\xi_{0:t}),\, P(\cdot|\tilde{\xi}_{0:t})\right) \leq \kappa_t \cdot \left\|\xi_{0:t} - \tilde{\xi}_{0:t}\right\|.$

Then the nested distance of the filtered spaces $\mathbb{P}_n = \left(\Xi, (\Sigma_t)_{t=0,\dots T},\, \hat{P}_n * k_{h_n}\right)$ equipped with the convolution measure $\hat{P}_n * k_{h_n}$ converges in probability to $\mathbb{P} = (\Xi, (\Sigma_t)_{t=0,\dots T}, P)$, i.e.,

$$P\left(\mathsf{dl}\left(\mathbb{P},\, \mathbb{P}_n\right) > \varepsilon\right) \to 0$$

as $n \to \infty$.

Generating trees

The tree generator algorithm we propose is based on a sample $(\xi_{i,1}, \dots, \xi_{i,T})$, $i = 1, \dots, n$. One replaces the probability distribution at the first stage $t = 1$ by the discrete measure $\sum_{i=1}^{b_t} p_i \delta_{\xi_{i,1}}$. This can be accomplished based on optimal quantizers, cf. Graf and Luschgy or by algorithms outlined in Pflug and Pichler (2014a, 2014b). Recursively, given that the tree is already established for t stages, each path $(\tilde{\xi}_1, \dots \tilde{\xi}_t)$ from the tree already constructed is being considered again. The conditional distribution is estimated from the samples by

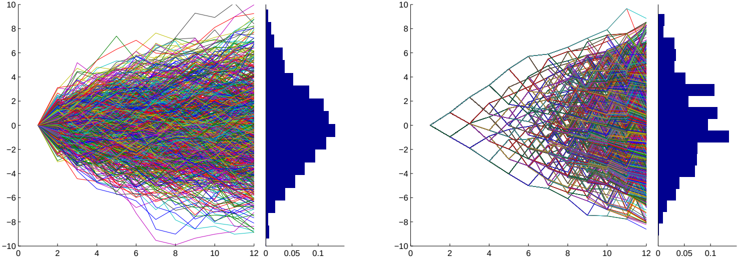
$$f(x_{t+1} | \, \tilde{\xi}_1, \dots \tilde{\xi}_t) \sim \hat{f}_n(x_{t+1} | \, \tilde{\xi}_1, \dots \tilde{\xi}_t),$$

This distribution is again approximated by a discrete probability measure. The parameters T (the desired height of the tree) and let $(b_1, \dots, b_T)$ the bushiness parameters per stage have to be chosen in advance.

The Algorithm

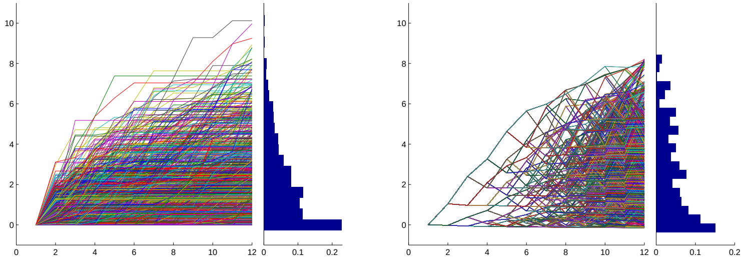
- **Determining the root.** The value of the process at the root is $\tilde{\xi}_0$ (deterministic). Its stage is 0. Set the root as the current open node.
- **Successor generation.** Enumerate the tree stagewise from the root to the leaves.
  1. Let ℓ be the node to be considered next and let $t < T$ be its stage. Let $\tilde{\xi}_0, \tilde{\xi}_1, \dots \tilde{\xi}_t$ be the already fixed values at node ℓ and all its predecessors. Find an approximation of the form $\sum_{i=1}^{b_t} p_i \delta_{x^{(i)}}$, which is close in the Wasserstein distance to the distribution with density
  $$f(x_{t+1} | \tilde{\xi}_0, \dots \tilde{\xi}_t) \sim \hat{f}_n(x_{t+1} | \tilde{\xi}_0, \dots \tilde{\xi}_t).$$
  2. Store the b_t successor nodes and assign to the tree the values $\tilde{\xi}(n_1) = x^{(1)}, \dots, \tilde{\xi}(n_{b_t}) = x^{(n_{b_t})}$ as well as their conditional probabilities $q(n_i) = p_i$ in the new tree.
- **Stopping Criterion.** If all nodes at stage $T - 1$ have been considered as parent nodes, the generation of the tree is finished.

Example: Markovian process



1000 sample paths from a Gaussian random walk and a binary tree of height 12 with 4095 nodes approximating it.

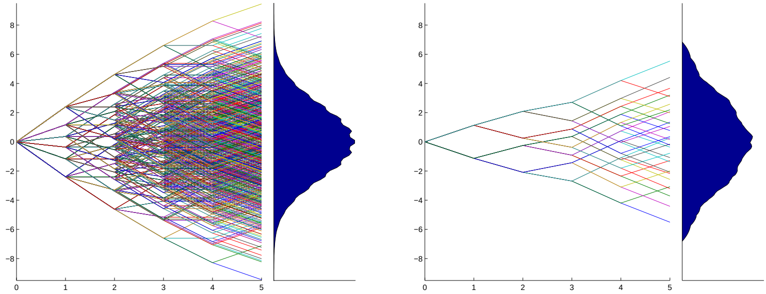
Examples: Non-Markovian process



1000 sample paths from a non-Markovian maximum process and a tree approximating it.

Example: Tree reduction

This example considers a tree as a starting process. The figure below left depicts a tree process with 1 237 nodes. Based on 10.000 samples we used our Algorithm to approximate the initial tree by a smaller one. Notice that a tree process does not have a density. Nevertheless, the algorithm still is able to approximate the initial tree and perform a tree reduction.



Model uncertainty: ambiguity

Traditionally, optimal decision making under uncertainty is done two steps:

- Step 1: Estimation of a probability model for the random scenarios
- Step 2: Finding the best decision given the estimated model

According to Ellsberg (1961) we face here two types of non-determinism:
Uncertainty: the probabilistic model is known, but the realizations of the random variables are unknown ("aleatoric uncertainty")
Ambiguity: the probability model itself is not fully known ("epistemic uncertainty").
Ambiguity sets $\mathcal{P}$: A family of probability models $\mathcal{P}$ which are all plausible models for the reality and we are uncertain about which concrete $P \in \mathcal{P}$ is the true one.

Problem formulation: Ambiguity

Let the basic problem be

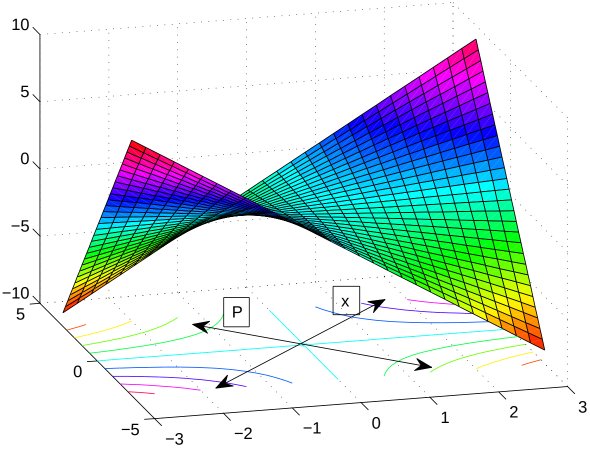
$$\min \left\{ \mathbb{E}_{\hat{P}}[Q(x,\xi)] \; : \; x \in \mathbb{X} \right\}$$

and let $\mathcal{P}$ be the ambiguity set. Then the ambiguity problem is

$$\min \left\{ \max \left\{ \mathbb{E}_{P}[Q(x,\xi)] \; : \; P \in \mathcal{P} \right\} \; : \; x \in \mathbb{X} \right\}.$$

Find the pair of optimal decision $x^* \in X$ which is good for all models $P \in \mathcal{P}$, among which there is a worst case model $P^* \in \mathcal{P}$.

The pair (x^*, P^*) forms a saddle point



A saddle point (symbolic picture)

Statistical estimation and confidence sets

- The ambiguity set $\mathcal{P}$ must reflect our current information about P
- If our information is based on statistical estimation, the ambiguity set must coincide with a confidence set
- by getting more or finer information, the ambiguity set may be reduced.

A simple portfolio optimization problem

$(\xi_1, \dots, \xi_M)$	random returns for M asset categories
$(x_1, \dots, x_M)$	portfolio weights
$Y_x = \sum_{m=1}^{M} x_m \xi_m$	portfolio return
$\mathcal{U}(Y_x)$	acceptability/utility functional

$$\left\| \begin{array}{l} \text{Maximize (in } x\text{)} : \ \mathcal{U}_P(Y_x) \\ \text{subject to} \\ x^\top \mathbb{1} = 1 \\ x \geq 0 \end{array} \right.$$

Equal weights is maximin for large ambiguity

With this insight, we may prove a remarkable result for distortion functionals:

$$\lim_{K\to\infty} \operatorname{argmax}_{\{\sum x_i=1, x_i\geq 0\}} \min_{\mathrm{d}_r(P,\hat{P})\leq K} \mathcal{U}_P(Y_x) = \frac{1}{M}\mathbb{1}.$$

Under large ambiguity, the optimal decision is the "equal weights" allocation.

The same result holds for the Markovitz model, if the distance is d_2.

Distortion utility functional: $\mathcal{U}(Y) = \int_0^1 F_Y(p)h(p)\,dp$

Average value-at-risk: $\mathbb{A}\mathrm{V@R}(Y) = \frac{1}{\alpha}\int_0^\alpha F_Y(p)\,dp$

Ambiguity in Copula

Suppose that an optimal portfolio problem, the marginal returns F_i are known and fixed and only the copula is unknown. We want to solve

$$\max_x \min_C \{\mathcal{U}(\sum_i x_i \xi_i) : \sum_i x_i = 1, \xi_i = F_i^{-1}(U_i), (U_1, \ldots, U_d) \sim C\}$$

where $\mathcal{U}$ is a comonotone additive utility functional (e.g. a distortion functional like the the $\mathbb{A}$V@R).
Then the minimax solution is

- C^* is the Fréchet upper bound (maxumal comonotonicity).
- x^* selects one single asset with has maximal individual utility

Completely different from the case with unknown marginals!

Multistage ambiguous models: balls in nested distance

As before, a baseline problem

$$\min \left\{ \mathcal{R}_{\hat{\mathbb{P}}}[Q\,(x,\xi)] \colon x \in \mathbb{X},\ x \lhd \mathfrak{F};\, \mathbb{P} = (\mathfrak{F}, P, \xi) \right\}$$

where the probability model is given by the nested distribution $\mathbb{P}$ is extended to the ambiguous model

$$\min_{x} \max_{\mathbb{P}} \left\{ \mathcal{R}_{\mathbb{P}}[Q(x,\xi)]\ :\ x \in \mathbb{X},\ x \lhd \hat{\mathfrak{F}};\, \mathbb{P} = (\hat{\mathfrak{F}}, P, \xi);\, \mathsf{dl}_r(\hat{\mathbb{P}},\mathbb{P}) \le \varepsilon \right\}.$$

$$\min \left\{ \max_{\mathbb{P} \in \mathcal{P}} \mathcal{R}_{\hat{\mathbb{P}}}[Q\,(x,\xi)] \colon x \in \mathbb{X},\ x \lhd \mathfrak{F};\, \mathbb{P} = (\mathfrak{F}, P, \xi) \right\}$$

In multistage models, we replace the Wasserstein distance by the nested distance dl for scenario trees and consider as ambiguity set the nested ball

$$\mathcal{P} = B_r(\hat{\mathbb{P}}, \varepsilon) = \left\{ \mathbb{P} :\ \mathbb{P} = (\hat{\mathfrak{F}}, P, \xi);\, \mathsf{dl}_r(\hat{\mathbb{P}},\mathbb{P}) \le \varepsilon \right\}.$$

The sequential algorithm

1. Set $n = 0$ and $\mathcal{P}_0 = \{\hat{\mathbb{P}}\}$ with $\hat{\mathbb{P}} \in \mathcal{P}$.
2. Solve the outer problem.

$$\min_x \max_{\mathbb{P}\in\mathcal{P}_n} \left\{\mathcal{R}_{\mathbb{P}}[Q(x,\xi)] \,:\, x \in \mathbb{X},\; x \lhd \hat{\mathfrak{F}};\, \mathbb{P} = (\hat{\mathfrak{F}}, P, \xi)\right\}.$$

   and call the solution x_n.

3. Solve the inner problem.

$$\max_{\mathbb{P}\in\mathcal{P}} \{\mathcal{R}_{\mathbb{P}}[Q(x_n,\xi)]\}$$

   to find the *worse case tree* $\mathbb{P}_{n+1}$. This can be accomplished by solving T linear problems, where T is the depth of the tree.

4. Set $\mathcal{P}_{n+1} = \mathcal{P}_n \cup \mathbb{P}_{n+1}$ and goto [2.] or stop.

Price of Ambiguity and Reward for Robustness

Let $\hat{\mathbb{P}}$ be the baseline model and let $x^*(\mathbb{P})$ be the optimal solution of the baseline problem. Likewise, let $\mathcal{P}$ be the ambiguity set and let $x^*(\mathcal{P})$ be the solution of the minimax problem. Under convex-concavity, the solution $x^*(\mathcal{P})$ of the minimax problem together with the worst case model $\mathbb{P}^*$ form a saddle point, meaning that the following inequality is valid for all feasible x and all $\mathbb{P} \in \mathcal{P}$

$$\mathbb{E}_{\mathbb{P}}[Q(x^*(\mathcal{P}), \xi)] \leq \mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}), \xi)] \leq \mathbb{E}_{\mathbb{P}^*}[Q(x, \xi)].$$

Let us call $\mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}), \xi)]$ the minimax value.

Define:

- The Price of Ambiguity.

$$\mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\mathcal{P}),\xi)] - \mathbb{E}_{\hat{\mathbb{P}}}[Q(x^*(\hat{\mathbb{P}}),\xi)] \geq 0.$$

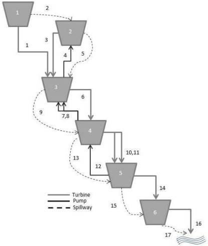
  "How much do I loose by implementing the minimax strategy $x^*(\mathcal{P})$ instead of the best strategy for the baseline model, if in fact the baseline model is true?"

- Reward for robust decisions.

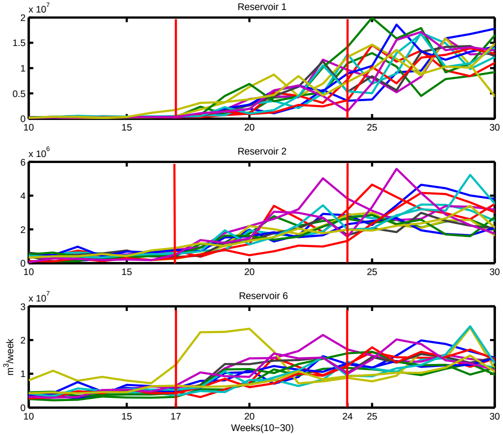
$$\mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathbb{P}),\xi)] - \mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}),\xi)] \geq 0.$$

  "How much do I gain, when I implement the minimax strategy $x^*(\mathcal{P})$ instead of the best strategy for the baseline model, if in fact the worst case model is true?"

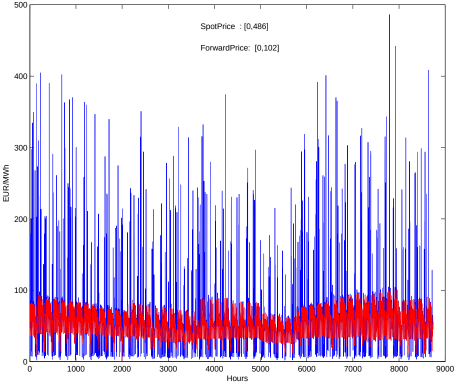
Management of a hydrosystem in the Austrian Alps



The scenario process consist of 5 components: Spot prices, Pumping prices, Inflows for 3 reservoirs. Statistical model selection methods were used to find that the inflows can be represented by a 3-dimensional $SARMA(1,2),(2,2)_{52}$ process, while the spot and pumping prices can be modeled by an independent process, a superposition of an additive error model based on forward prices and a spike generating process.



Observations for Inflows



Observations for Spot/Forward prices

The decision model

maximize

$\lambda \mathbb{E}[x^c_T] - (1-\lambda)\mathbb{A}\mathrm{V@R}_{1-\alpha}[-x^c_T]$

subject to

$0 \leq x^f_{t,i} \leq \overline{x}^f_i,$

$\underline{x}^s_j \leq x^s_{t,j} \leq \overline{x}^s_j,$

$x^{,s}_{end,j} \leq x^s_{T,j},$

$x^s_{t,j} = x^s_{t-1,j} + \xi^f_{t,j} + \sum_{\{i \in I | P_{max} > 0\}} A_{i,j} \cdot x^f_{t-1,i} + \sum_{\{i \in I | P_{max} = 0\}} A_{i,j} \cdot x^f_{t,i},$

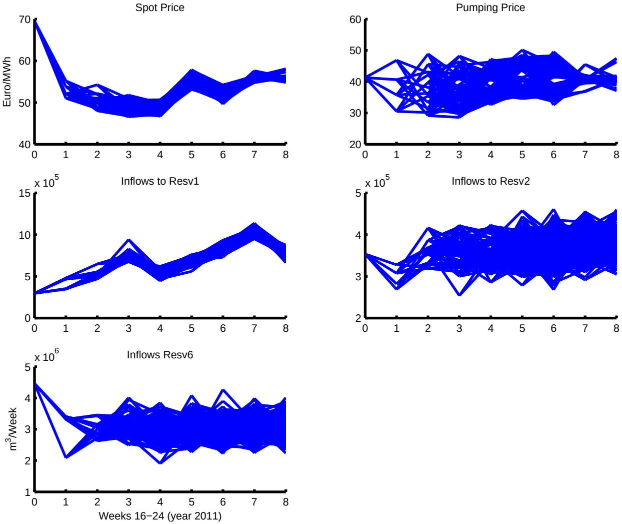
$x^e_{t,i} = x^f_{t-1,i} \cdot k^i \cdot \triangle t_{(t-1)},$

$x^c_t = x^c_{t-1} \cdot (1+r)^{\triangle t_{(t-1)}} + \sum_{\{i \in I | k^i > 0\}} x^e_{t-1,i} \cdot \xi^e_t + \sum_{\{i \in I | k^i < 0\}} x^i_{t-1,i} \cdot \xi^p_t.$

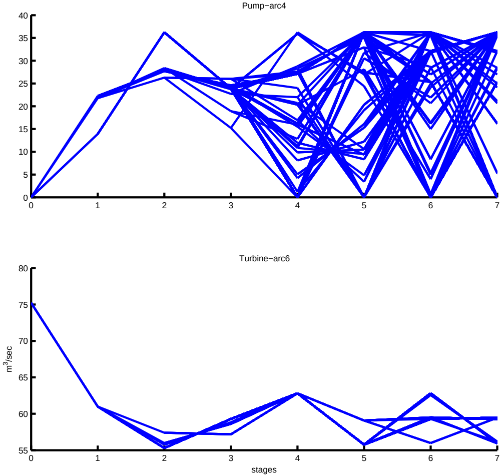
Generating a scenario tree

We generate a scenario tree in a way that the nested distance between the scenario process and the scenario tree is as small as possible.

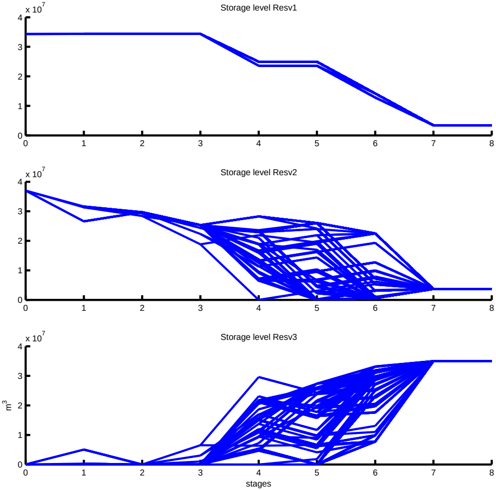
Number of stages	8
Minimal bushiness per stage	2,2,2,1,1,1,1,1
Maximal distance per stage	5,5,5,7,7,7,10,10
Number of scenarios (leaves)	392
Number of nodes 1532	



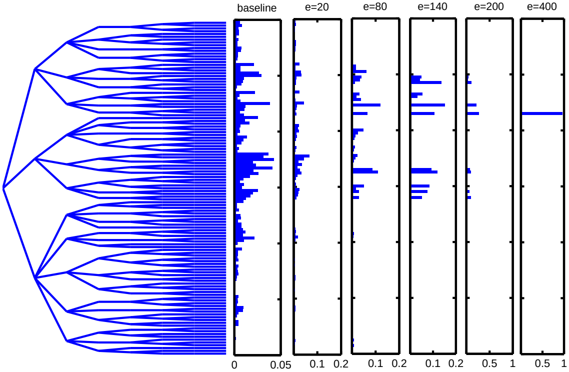
The generated five-dimensional tree



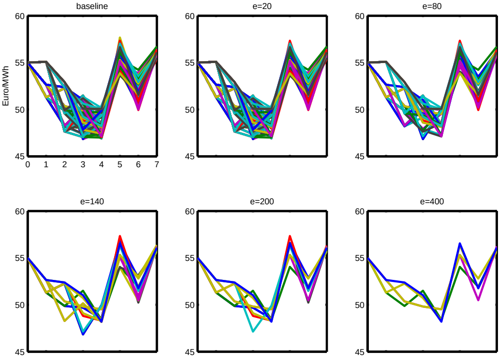
The pumping (top) and turbining (bottom) decisions



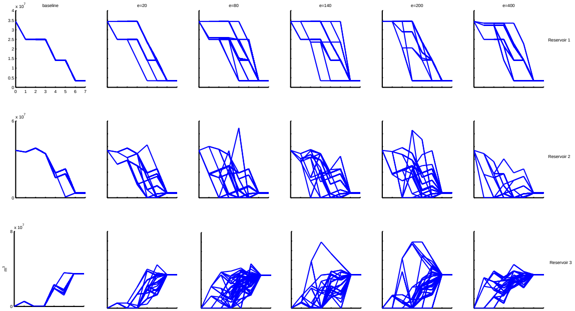
The storage levels



The typical picture: The larger is the ambiguity radius, the simpler is the worst case tree.



The worst case spotprice trees: Only the arcs with a minimum probability are shown.



The minimax decisions: They get more complicated with increasing ambiguity radius: Decisions lying on bounds are avoided.

Price of ambiguity: 2.3%.

Reward for robustness: 7.5%.

Conclusions

- Multistage stochastic programs need approximation techniques. We showed how the approximation error can be controlled.
- The way from data to models can be parametric or nonparametric
- In order to capture both: scenario uncertainty (aleatoric uncertainty) and probability ambiguity (epistemic uncertainty) we use a probabilistic maximin approach
- The ambiguity neighborhood should be chosen in such a way that it corresponds to statistical confidence regions for which bounds for the covering probability are available.
- Bounds may be used if instead of approximations, if the original problems is quite difficult to solve.

Lower and upper bounds

Often the problem is computationally such complex, that it is advisable to find quick bounds than to calculate the exact solution. If a feasible solution is found, for which the objective value can only be 1% (say) larger than a valid lower bound, then one may stop and avoid cumbersome calculations for a possible minor effect. The general principle of bounding is the following:
Lower bounds can be found by relaxation of constraints or by finding minorants of the objective function (e.g. by Jensen's inequality)
Upper bounds can be found by inserting feasible solutions or by finding majorants of the objective function (e.g. by Edmundson-Madansky inequality).

Lower bounds by changing the probability measure

These classes of lower bounds are based on the following observation:
If the functional $P \longmapsto \mathcal{R}_P(\cdot)$ is concave (i.e. the mapping $P \mapsto \mathcal{R}_P(Y)$ is concave for all random variables Y for which $\mathcal{R}$ is defined), then the mapping $P \mapsto v^*(P)$ is also concave.
Consequently, if $P = \sum_i w_i P_i$, then

$$\sum_i w_i v^*(P_i) \leq v^*(P).$$

Refinement Chains

While the dissection of P into its atoms (δ_{ω_i}) is the most extreme dissection, one may also dissect it into a convex combination of probabilities sitting on two ore more, but not all leaves. We may consider dissections of P into a convex combination of probabilities, all of them sitting on j scenarios. These probabilities need not to sit on disjoint sets, to the contrary, all of them may contain f fixed scenarios.

If a refining sequence of dissections of P can be found, then a monotonic sequence of lower bounds can be found.

A refinement chain is of the structure

$$\begin{array}{c} \Omega \\ (\Omega_1^{(\ell)}, \Omega_2^{(\ell)}, \ldots, \Omega_{m_\ell}^{(\ell)}) \\ \vdots \\ (\Omega_1^{(2)}, \Omega_2^{(2)}, \ldots, \Omega_{m_2}^{(2)}) \\ (\{\omega_1\}, \{\omega_2\}, \ldots, \{\omega_k\}) \end{array}$$

where each row is a collection of subsets of the probability space Ω with the property that their union covers the whole space $\Omega = \cup_i \Omega_i^{(j)}$ for all j and that each set $\Omega_i^{(j+1)}$is the union of sets from the next more refined collection

$$\Omega_i^{(j)} = \cup_{\Omega_s^{(j-1)} \subseteq \Omega_i^{(j)}} \Omega_s^{(j-1)}.$$

To a refinement chain of the probability space Ω there corresponds a chain of dissections of the probability P into probability measures $P_i^{(j)}$

$$\begin{array}{c} P \\ (P_1^{(\ell)}, \ldots P_{m_\ell}^{(\ell)}) \\ \vdots \\ (P_1^{(2)}, \ldots, P_{m_2}^{(2)}) \\ (P_1^{(1)} = \delta_{\omega_1}, \ldots, P_k^{(1)} = \delta_{\omega_k}) \end{array} \tag{3}$$

such that

(i) $P_i^{(j)}$ has support $\Omega_i^{(j)}$

(ii) P can be written as $P = \sum_{i=1}^{m_j} \pi_i^{(j)} P_i^{(j)}$

(iii) each $P_i^{(j)}$can be written as a convex combination of probabilities from the refined collection $\left\{ P_i^{(j-1)} \right\}$.

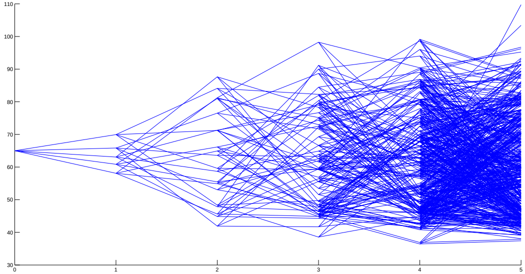
It is evident that given such a refinement chain leads to a chain of lower bounds.

Case study: a multistage inventory problem

Random demands have to be satisfied from an inventory. If the demand exceeds the stock, it will be satisfied by rapid orders from a different source, which come at a higher price. At each time step (stage), orders can be placed, and they will be delivered one period later. The objective is to minimize the expected disutility of the total costs where profits are considered as negative costs. Demands are the only random quantities in the model, all financial quantities are assumed to be already discounted to the present.

Computational results

We consider the inventory model using the random demands s modeled as the scenario tree below. It has 540 scenarios and 806 nodes. The full stochastic problem is composed by 4304 scalar variables and 2693 constraints.

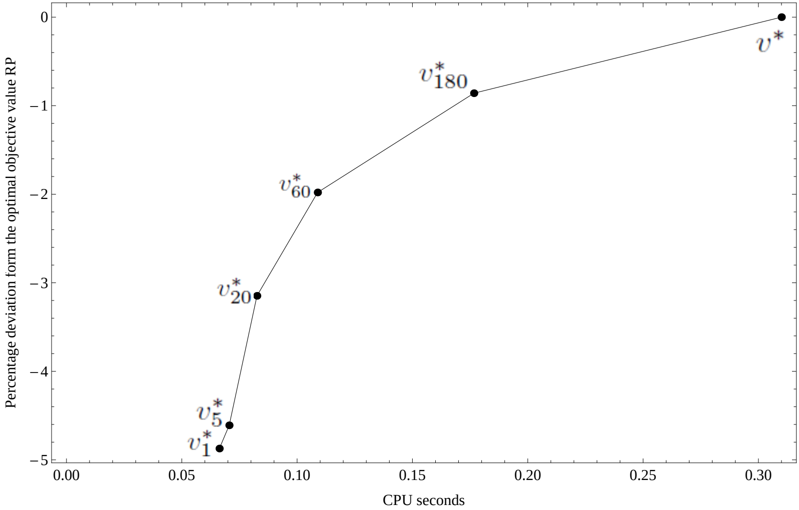


Scenario tree structure generated on the basis of an time-inhomogeneous exponential auto-regressive $AR(1)$ model, with demand values ξ_t, $t = 0, \ldots, 5$ represented on the y-axis.

Refinement chain with disjoint subsets ($f = 0$)

j	# subproblems s	Objective v.	CPU s. per subpr.
1	540	$v_1^* = -2198.81$	0.066
5	108	$v_5^* = -2193.31$	0.0705
20	27	$v_{20}^* = -2162.65$	0.0825
60	9	$v_{60}^* = -2138.17$	0.108
180	3	$v_{180}^* = -2114.67$	0.176
540	1	$v^* = -2096.66$	0.31

Percentage deviation form the optimal value v^*

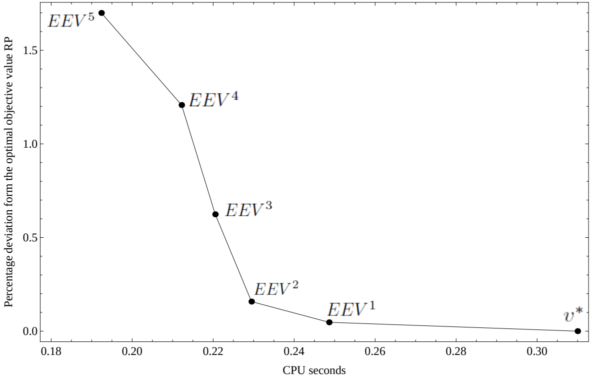


Upper bounds

Comparison between the stochastic, the expected value solution and upper bounds by inserting (sub)solutions for the multistage inventory problem ($k = 540$ scenarios).

problem type	x_0	Objective v.	CPU s.
$v^*(\mathbb{P}_{\mathbb{E}(\xi)})$ (lower b.)	61.4188	-2199.70	0.01
$v^*(\mathbb{P})$	63.8405	-2096.66	0.31
$EEV^1 = v^*(\mathbb{P}, \bar{x}_{0:0})$	61.4188	-2095.67	0.10
$EEV^2 = v^*(\mathbb{P}, \bar{x}_{0:1})$	61.4188	-2093.35	0.12
$EEV^3 = v^*(\mathbb{P}, \bar{x}_{0:2})$	61.4188	-2083.58	0.29
$EEV^4 = v^*(\mathbb{P}, \bar{x}_{0:3})$	61.4188	-2071.34	0.26
$EEV^5 = v^*(\mathbb{P}, \bar{x}_{0:4})$	61.4188	-2061.03	0.28

Percentage deviation



The upper approximation $\bar{P}$

A very large problem

We generated a tree with 725760 scenarios and 1262417 nodes. This problem could not be solved by our solver, but the lower bound approximation was possible.

j	# subprob.	Objective v.	CPUs/subprob.	total CPUs.
8	90720	$v_{8}^{*} = -3967.61$	0.0287	2612.29
56	12960	$v_{56}^{*} = -3941.06$	0.0539	698.71
336	2160	$v_{336}^{*} = -3921.15$	0.1990	430.011
2016	360	$v_{2016}^{*} = -3902.77$	0.9385	337.89
10080	72	$v_{10080}^{*} = -3884.23$	5.4852	394.94
40320	18	$v_{40320}^{*} = -3869.42$	27.4512	494.123
120960	6	$v_{120960}^{*} = -3857.06$	166.5516	999.31
362880	2	$v_{362880}^{*} = -3842.61$	1274.424	2548.848