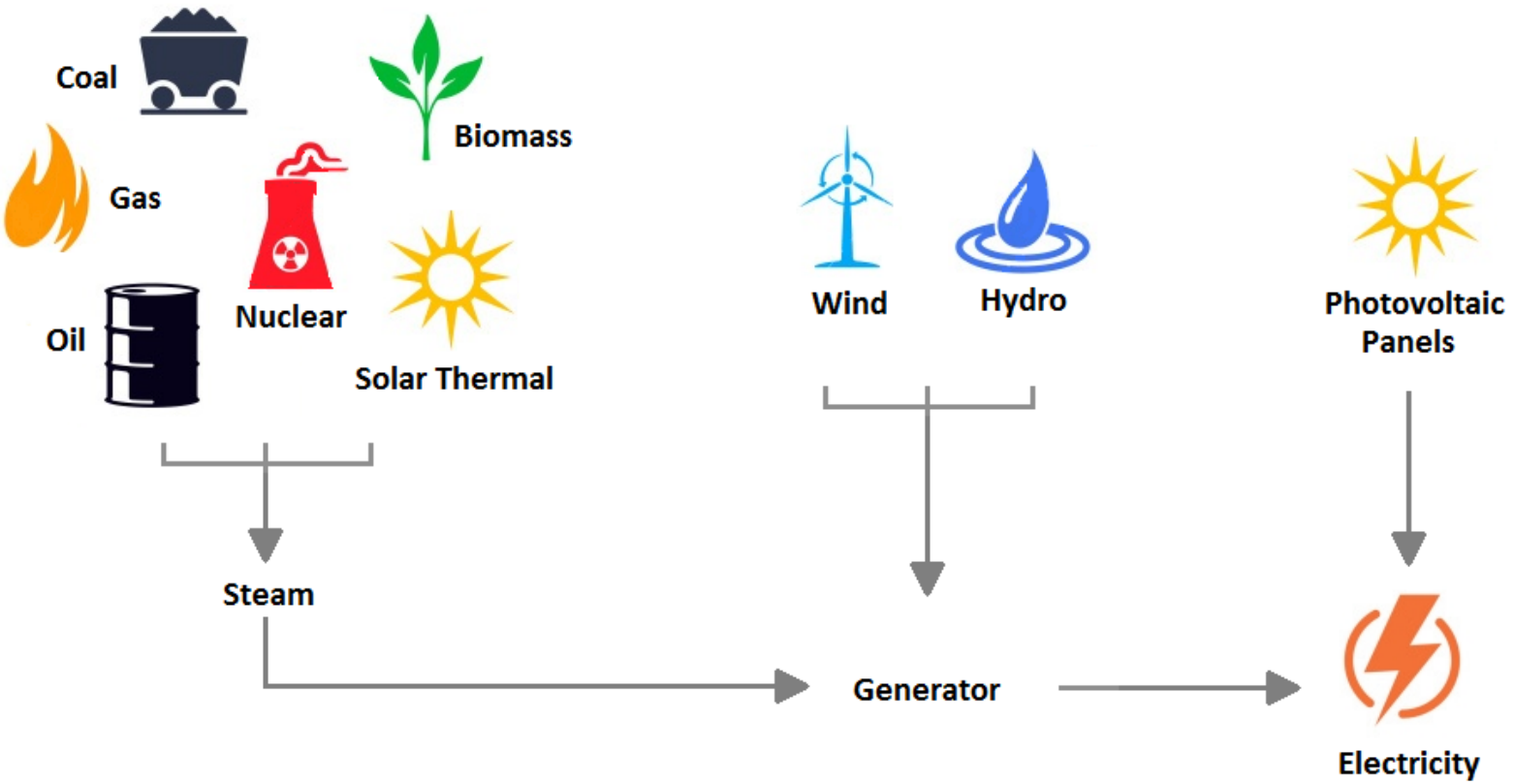
Part I: Decision making under uncertainty

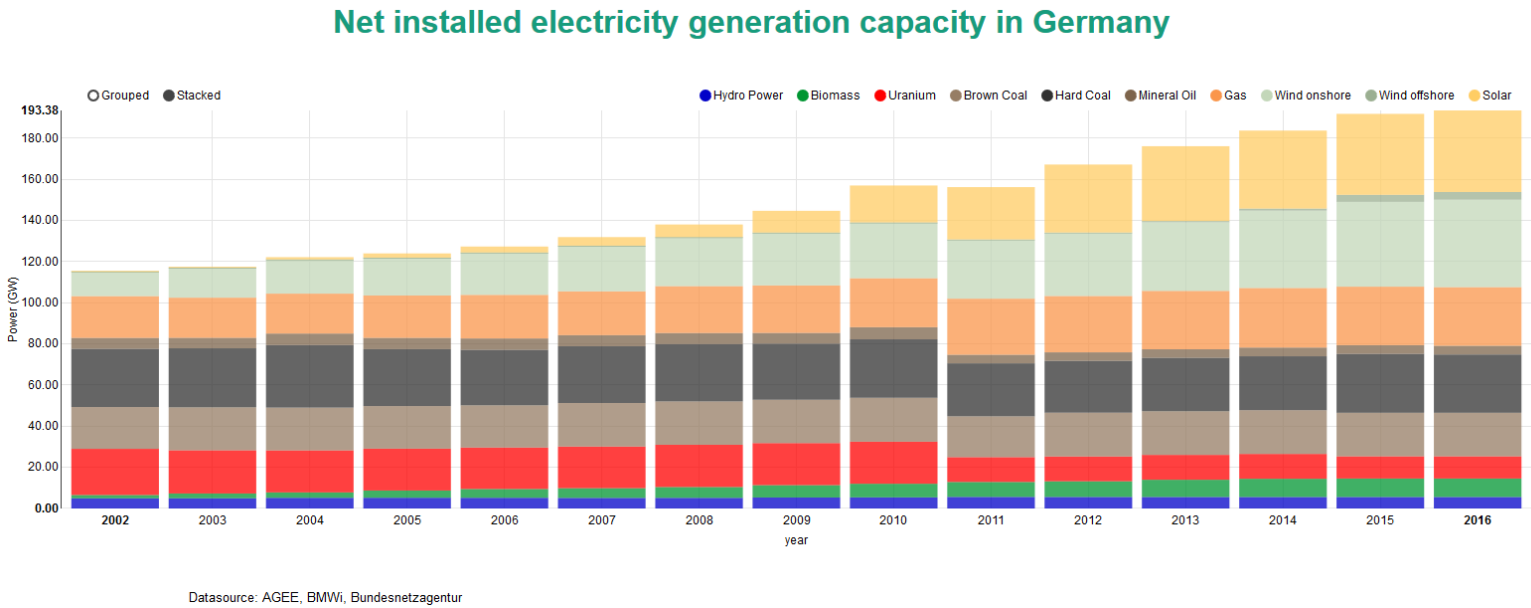
Georg Ch. Pflug
EES-UETP

July 3, 2016

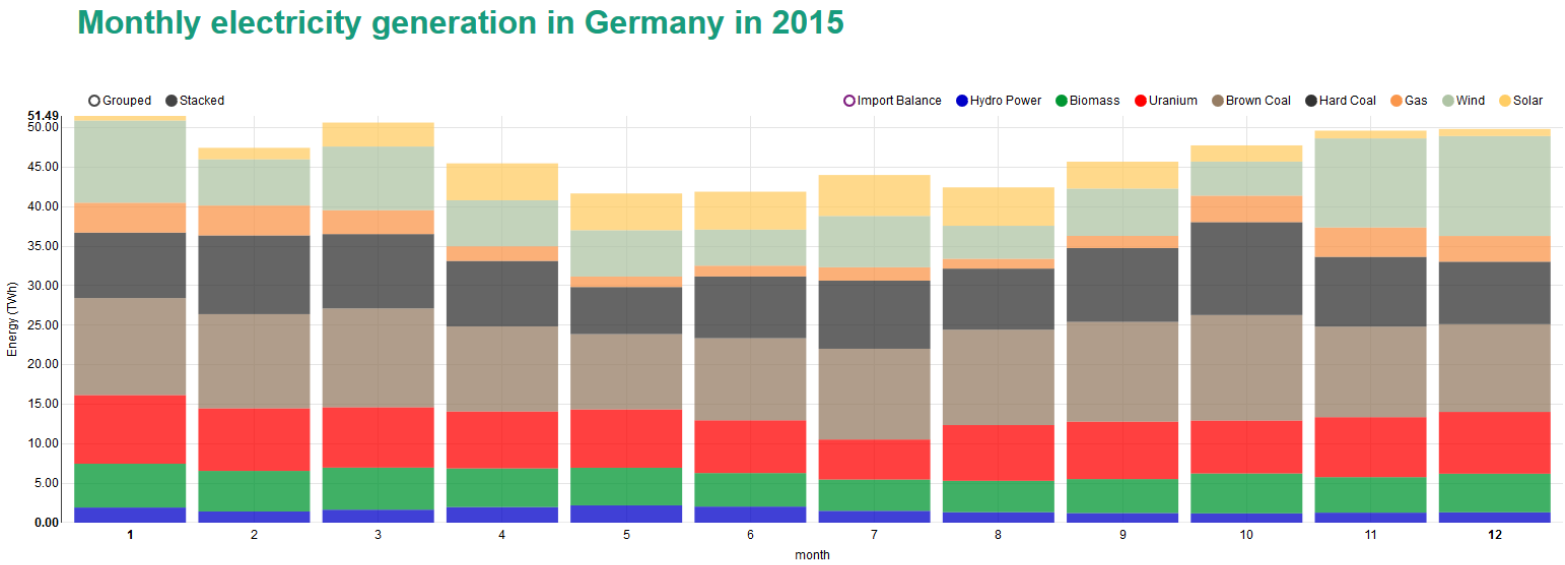
Electricity production



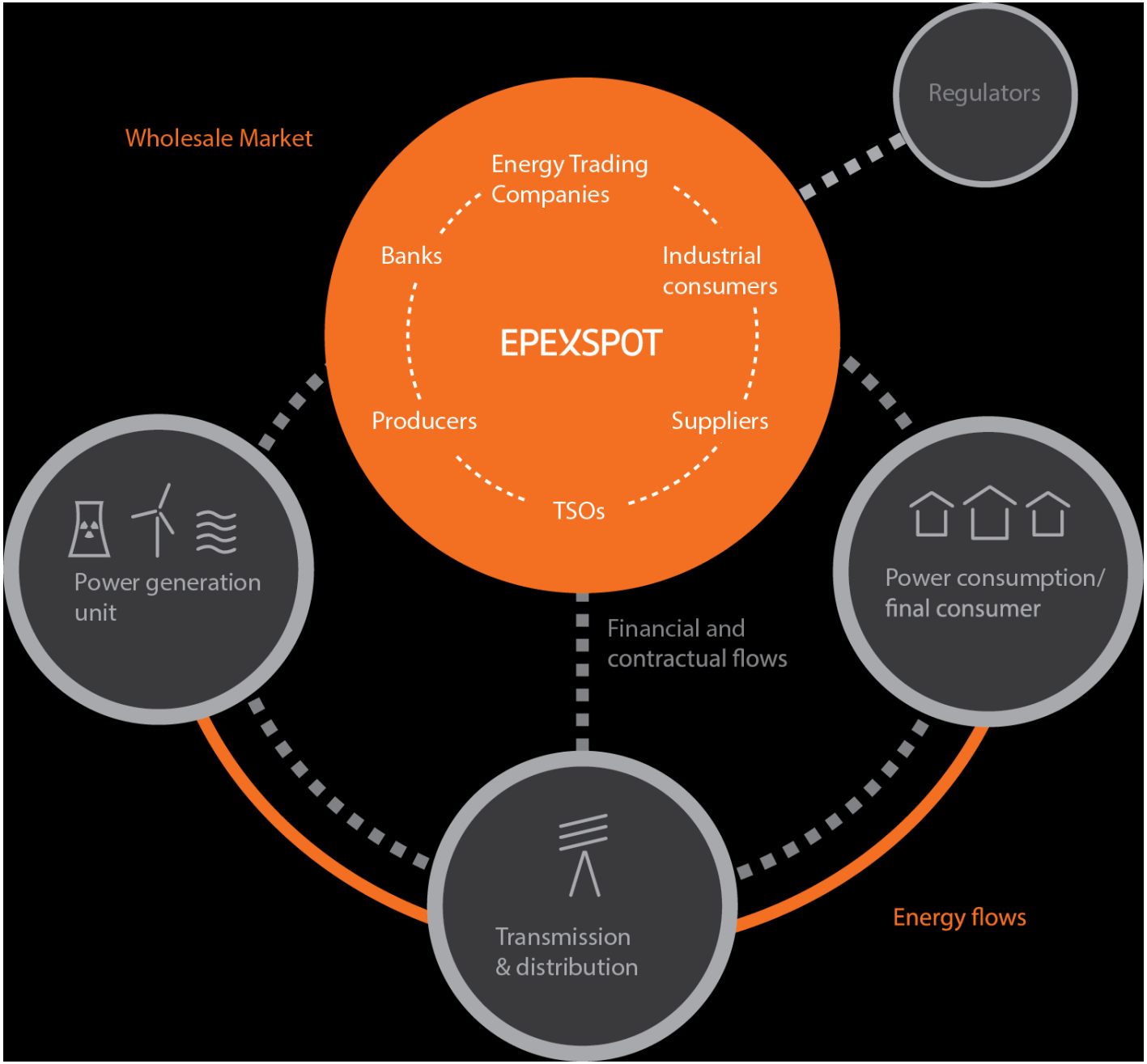
Production mix/installed



Production mix/used



Datasource: 50 Hertz, Amprion, Tennet, TransnetBW, EEX


Wholesale Market
Energy Trading Companies
Banks
Industrial consumers
EPEXSPOT
Producers
Suppliers
TSOs
Regulators
Power generation unit
Power consumption/ final consumer
Financial and contractual flows
Transmission & distribution
Energy flows

Risk factors

Decision making in electricity management depends on many uncertainties and risks.

- Uncertainty: A factor which is unknown and can be modelled by a random variable or random process
- Risk: An uncertainty, which may lead to unwanted situations

Risk factors:

- Resource costs for non-renewables
- Availability of renewables
- prices for Allowables (CO2-certificates)
- Demands
- Prices for short term delivery contracts (spot market)
- Prices for long term delivery constracts (futures)

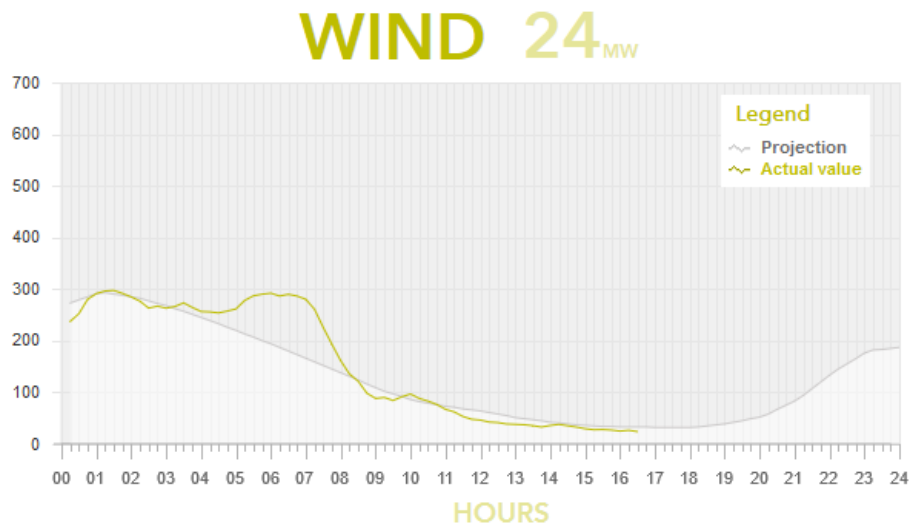
Gas prices

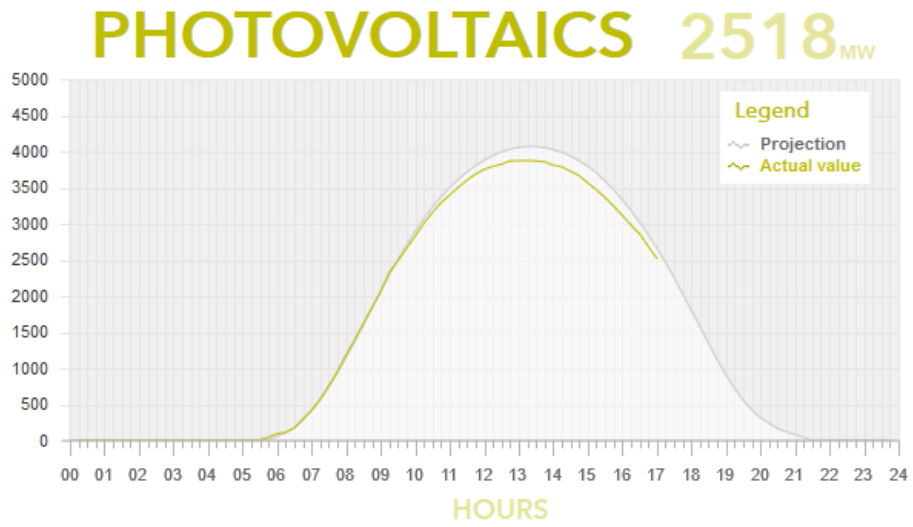

Daily Reference Price
Price
30
25
20
15
10
Jan '12
Jun '12
Dec '12
Jun '13
Dec '13
Jun '14
Dec '14
Jun '15
Dec '15
Jun '16

Coal prices

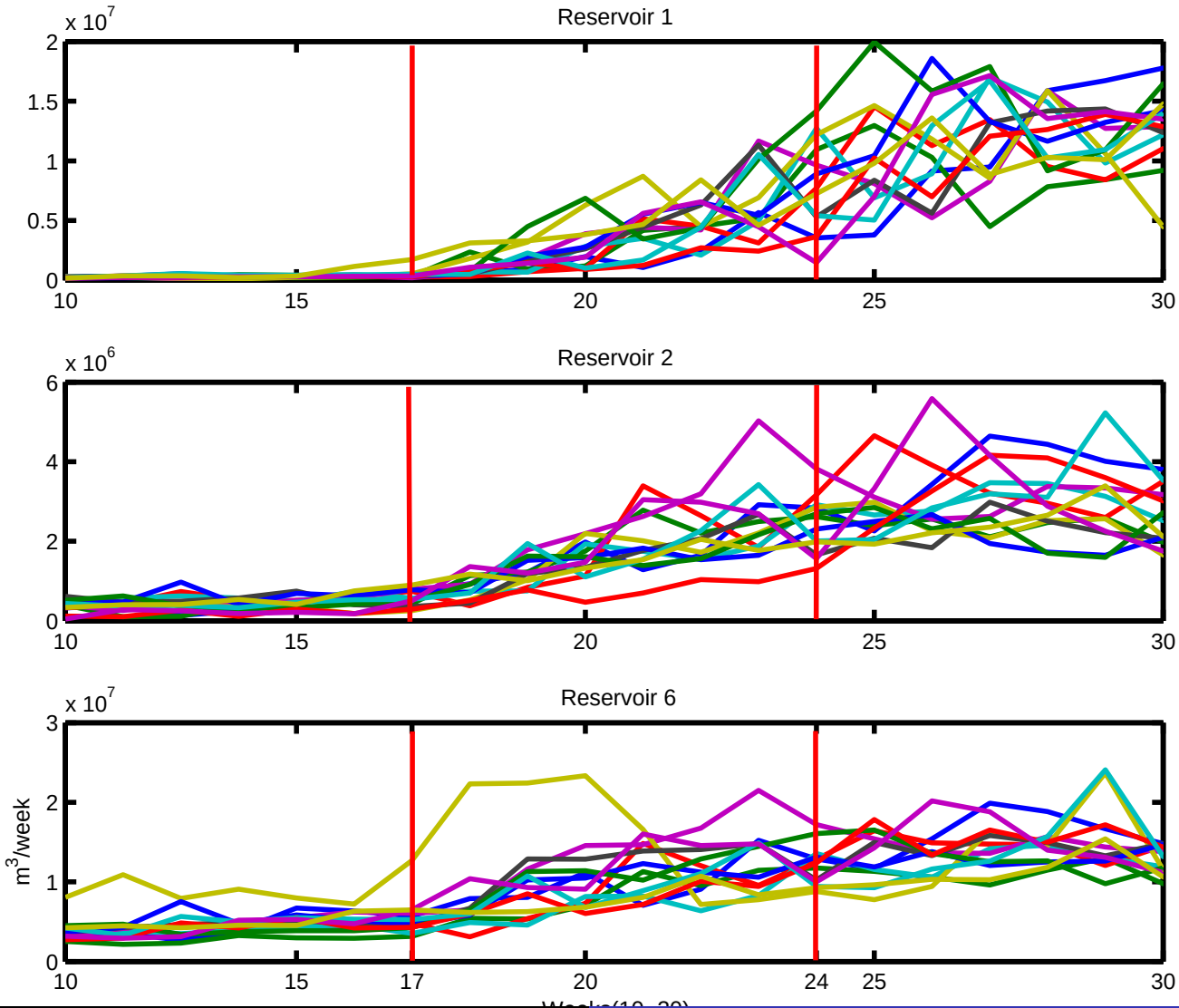

52.5
50
47.5
45
42.5
Price
11. Apr
25. Apr
9. May
23. May
6. Jun
20. Jun
Settlement Price

Renwables-availability

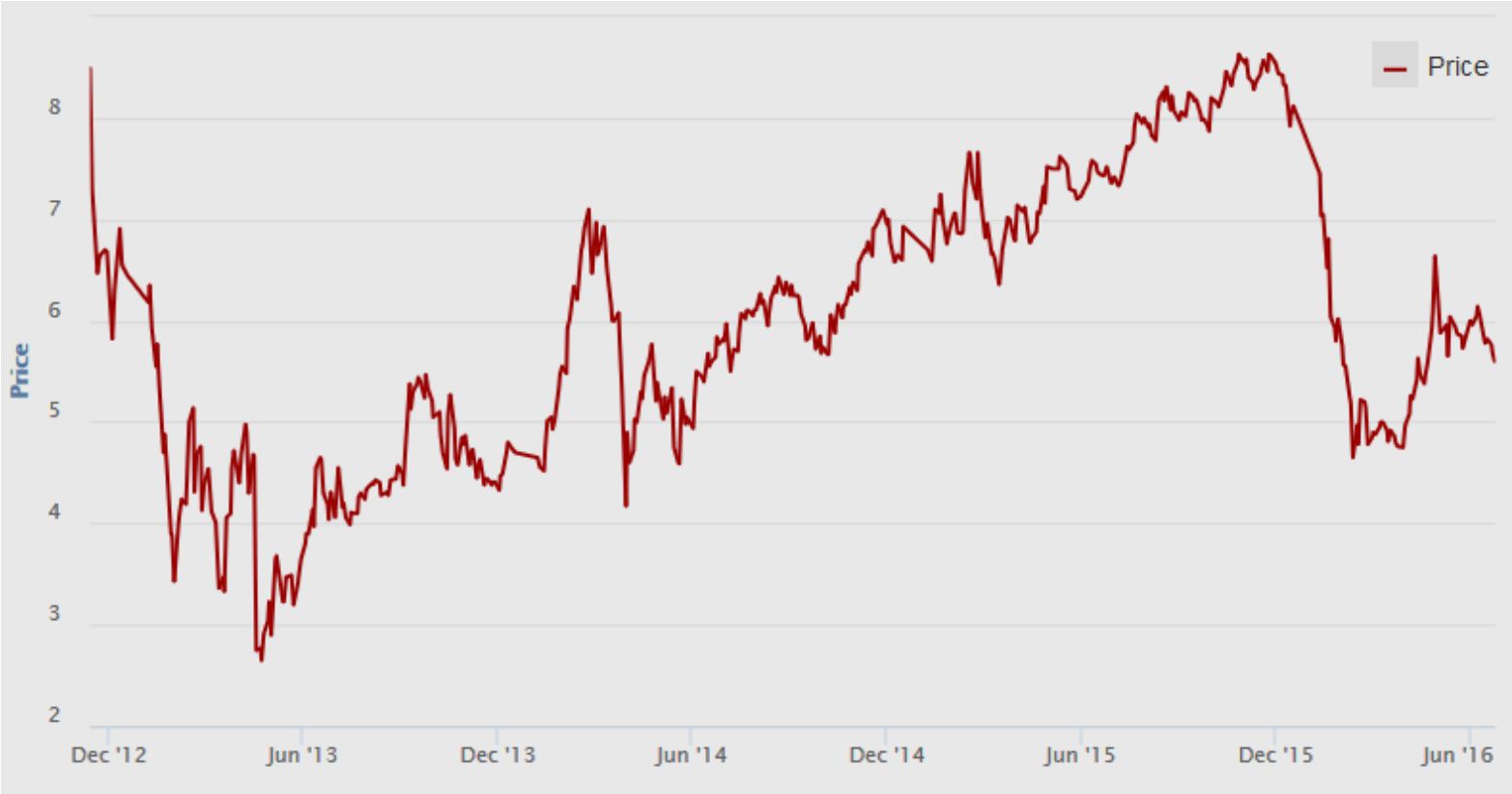




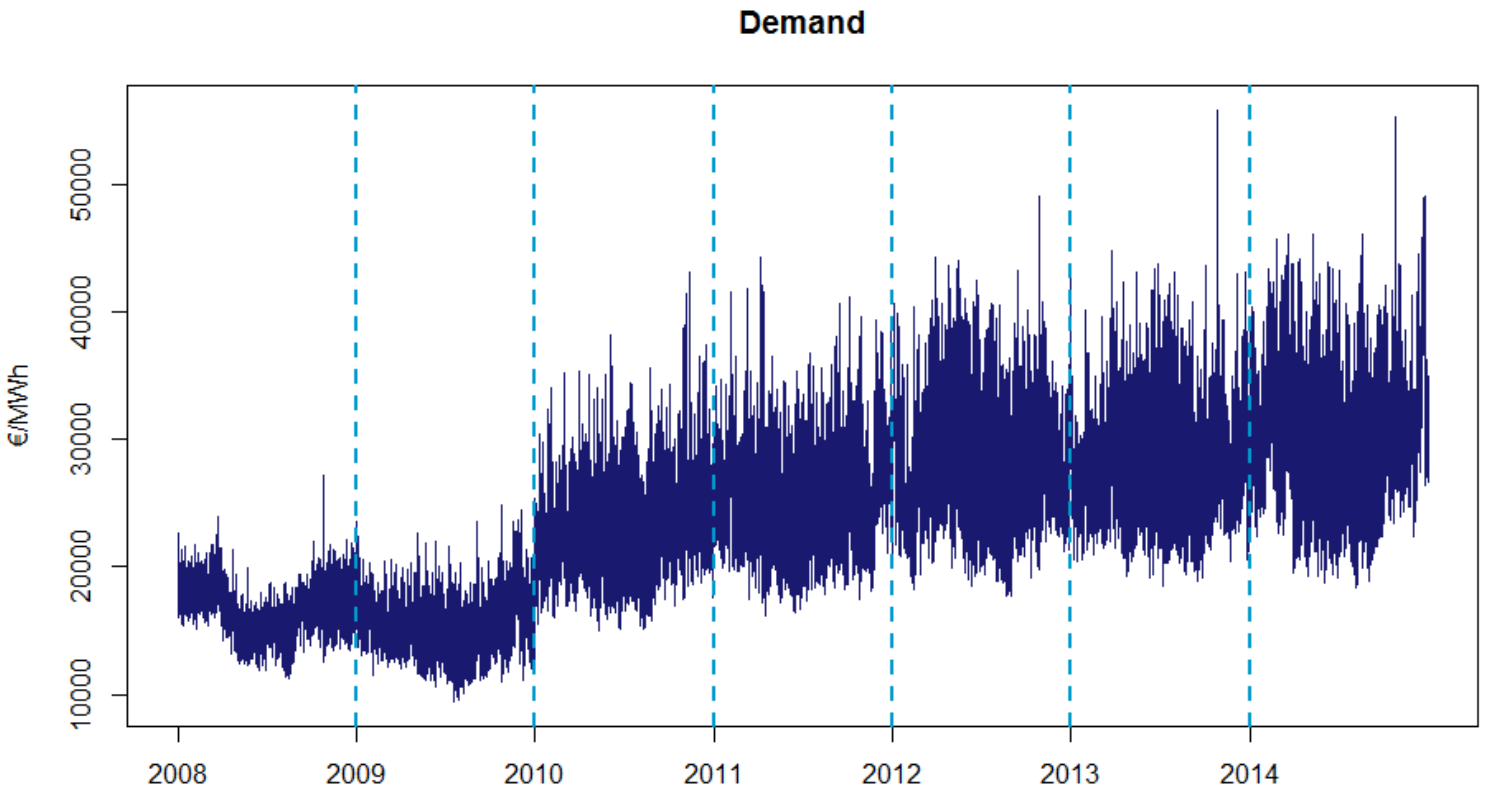
Water inflows



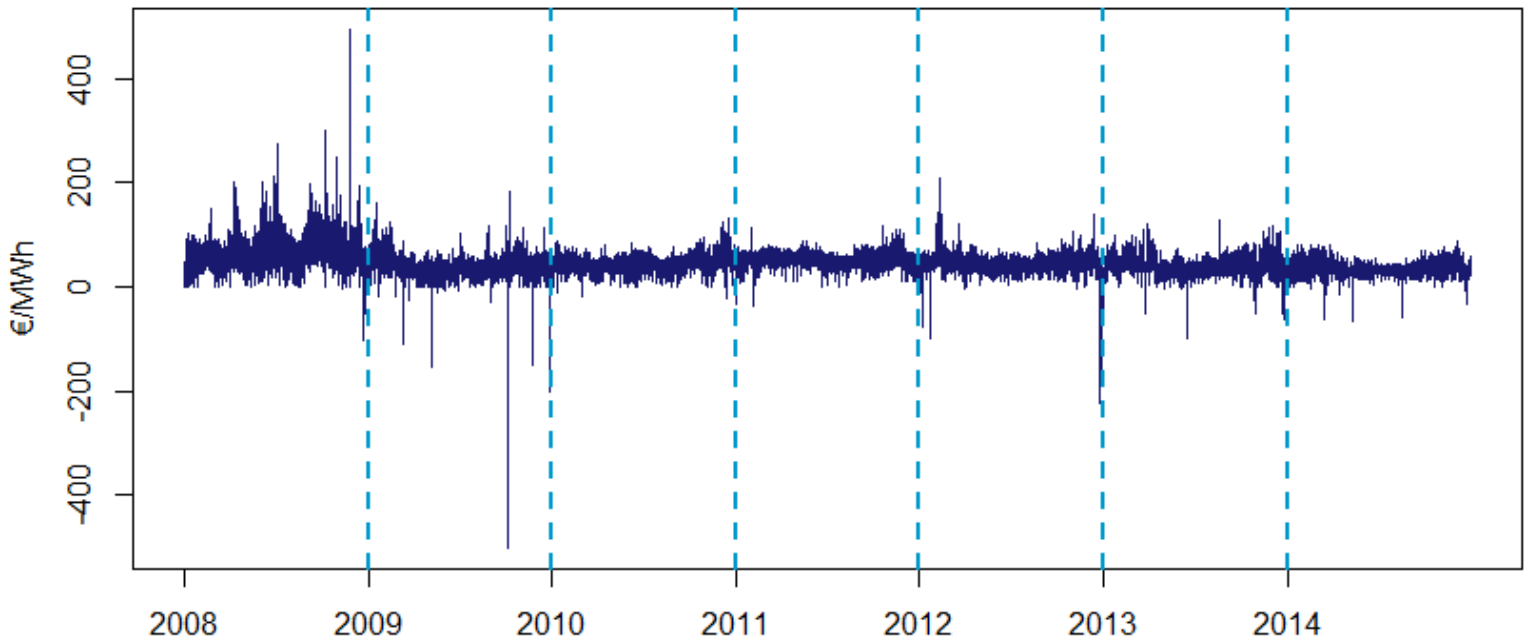
CO2-certificate prices



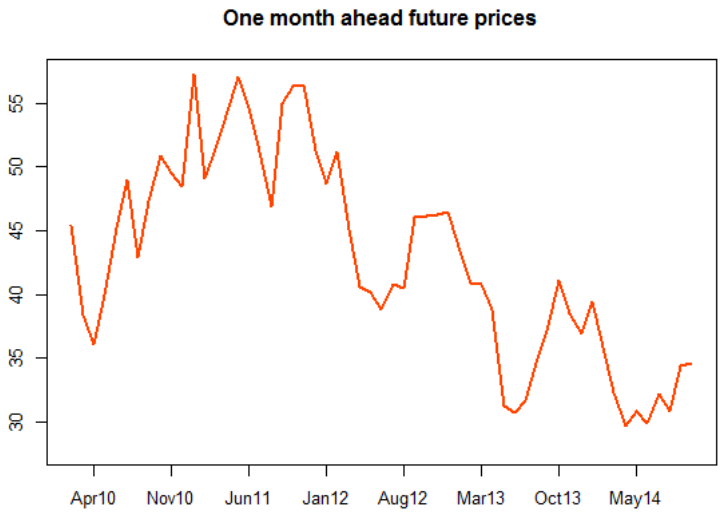
Demands

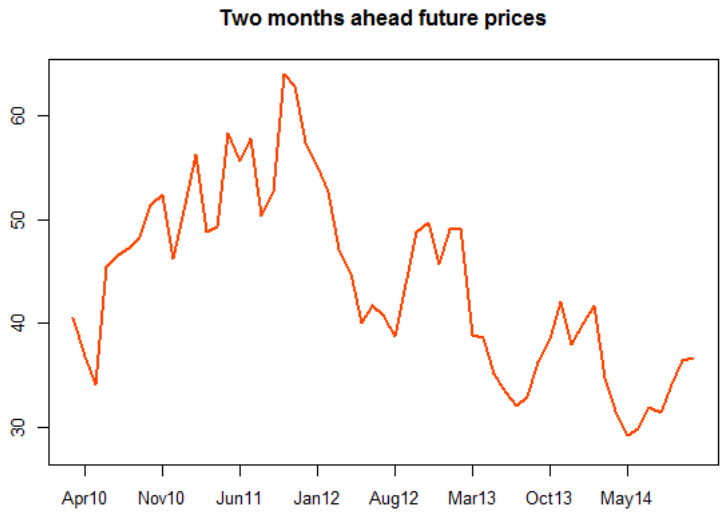

Demand
€/MWh
10000
20000
30000
40000
50000
2008
2009
2010
2011
2012
2013
2014

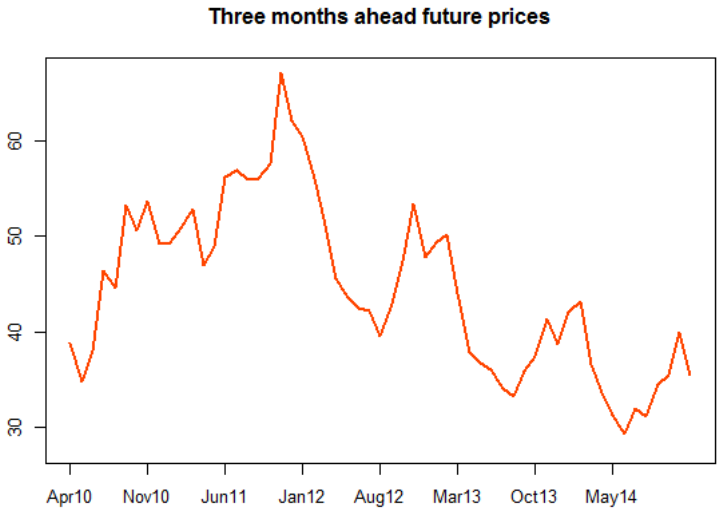
Spot market prices



Future prices







Stochastic optimization

Stochastic optimization is the technique for making optimal decisions under uncertainty and risk. According to Ellsberg (1961) we distingusih between

- Aleatoric uncertainty: the probabilistic model is known, but the realizations of the random variables are unknown. A possible realization is called a *scenario* (*scenario vector*, *scenario process*, *scenario tree*).
- Epistemic uncertainty: the probability model itself is not fully known ("Ambiguity"). A possible choice of the probability model is called a *scenario model*.

We use methods of stochastic optimization mainly for three tasks:

- **Pricing of contracts:** We have to find the lowest price, which - considering all hedging (risk reducing) strategies - is acceptable for the contract seller, or the maximal price, which is acceptable for the buyer. This includes bidding in electricity markets.
- **Optimal production and trading strategies:** We have to find the optimal strategies for managing a production and trading portfolio.
- (Real options: We have to find the optimal yes-no decisions for investment plans-will not be treated in this lecture).

The lecture plan

- Stochastic optimization
- Measuring risk: Risk functionals
- Modeling risk: Scenario models
- Pricing of electricity contracts: Single level and bilevel approaches
- Managing production and trading risks: baseline models and ambiguity models

The basic formulation of a stochastic optimization problem

$$\min\{\mathcal{R}[Q(x_0, \xi_1, x_1)] : x \in \mathbb{X}\},$$

where

$\mathcal{R}$	a risk functional
ξ	a random variable or random vector
x_1, x_2	the decisions,
$Q(x_0, \xi_1, x_1)$	the cost function,

Some nomenclature

- If only one decision has to be made in the beginning, this is called a single-stage problem
- If one decision is to be made right away and a recourse decision can be made after observing ξ, this is a two-stage problem. A two stage problem is typically formulated as

$$\min\{Q_0(x_0) + \min \mathcal{R}[Q_1(\xi, x_1)] : x_1 \in \mathbb{X}_1(x_0)\} : x_0 \in \mathbb{X}_0\}$$

  Here Q_0 is called the first stage costs and Q_1 is the recourse function.
- If $\mathcal{R}$ is the expectation, the problem is called risk-neutral, otherwise it is called risk-sensitive or risk-adverse

Incorporating risk

- *Risk-neutral* The optimization maximizes the expected profit or minimizes the expected costs.
- *Risk-averse* The optimization contains nonlinear functionals either in the cost or profit function or in the probability
  - Utility fnctionals (nonlinear functionals in the profit)
  - Risk functionals (nonlinear functionals in the costs)

Let Y be a profit/loss (P& L) variable defined on a probabiliy space $(\Omega, \mathcal{F}, P)$. The pertaining cost variable is $-Y$.
A risk functionals is

$$\mathcal{R}_P(-Y).$$

An utility functional is

$$\mathcal{U}_P(Y) = -\mathcal{R}_P(Y).$$

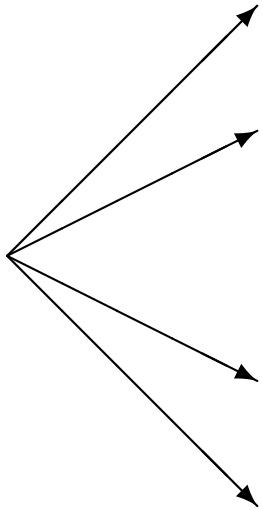
The basic formulation of a multistage stochastic program

$$\min\{\mathcal{R}[Q(x_0, \xi_1, \ldots, x_{T-1}, \xi_T)] : x \lhd \mathcal{F}\},$$

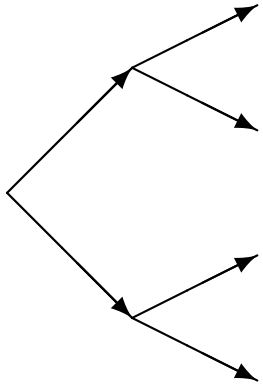
where

$\xi = (\xi_1, \ldots, \xi_T)$	a random scenario process defined on $(\Omega, \mathcal{F}, P)$
$x = (x_0, \ldots, x_{T-1})$	the sequence of decisions,
$Q(x_0, \xi_1, \ldots, \xi_T)$	the cost function,
$\mathcal{F} = (\mathcal{F}_1, \ldots \mathcal{F}_T)$	a filtration (an increasing sequence of σ-algebras),
$\xi \lhd \mathcal{F}$	ξ is adapted to $\mathcal{F}$, i.e. $\sigma(\xi) \subseteq \mathcal{F}$
$x \lhd \mathcal{F}$	the nonanticipativity condition.

Single-,two- and multistage

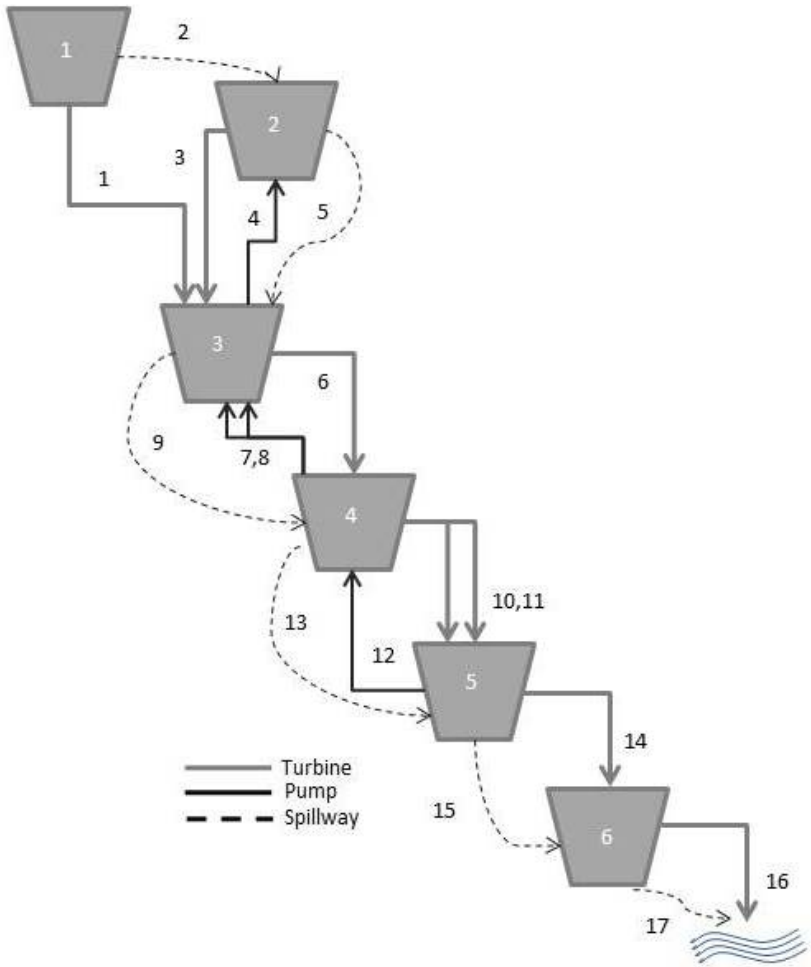


Single- or twostage



Multistage

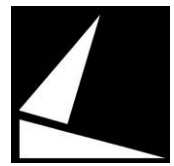
Example: Hydrostorage optimization



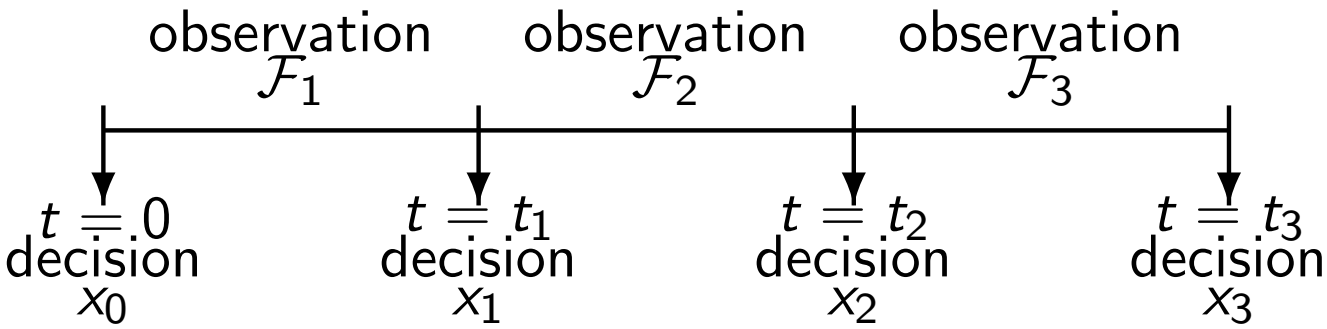
For each time step t the volume x_t to be turbined is determined, while the market price for energy η as well as the inflows ξ to the reservoir are observed only one period later. The reservoir balance is

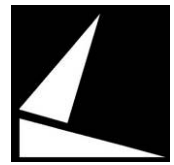
$$\begin{aligned} V_{t+1} &\leq V_t - x_t + \xi_{t+1} \\ V_{t+1} &\leq V_{max} \end{aligned}$$

The links between the decision stages are formulated as constraints.
A myopic model (single- or two-stage) is not appropriate here.

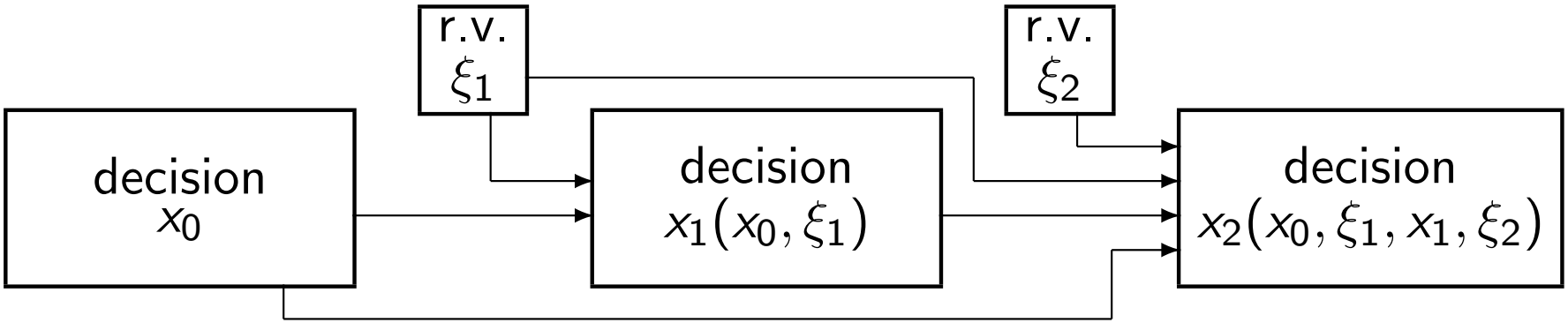


Non-anticipativity





Multistage stochastic decision processes



Linear stochastic recourse problems

$$\min \{c_0(x_0) + \mathbb{E}_{\xi_1} [\min\ c_1(x_1, \xi_1)] : \quad (x_0, x_1) \in \mathbb{X}\}$$

where the feasible set $\mathbb{X}$ is given by

$$\begin{aligned} W_0 x_0 &\geq h_0 \\ A_1\, x_0 + W_1\, x_1 &\geq h_1(\xi_1) \end{aligned}$$

Linear stochastic multistage problems

$$\min \{c_0(x_0) + \mathbb{E}_{\xi_1}[\min\, c_1(x_1, \xi_1) + \mathbb{E}_{\xi_2}[\min\, c_2(x_2, \xi_2) + \\ + \mathbb{E}_{\xi_T}[\cdots + \mathbb{E}_{\xi_T}[\min\, c_T(x_T, \xi_T)]\ldots]]] : x \in \mathbb{X}\},$$

where the feasible set $\mathbb{X}$ is given by

$$\begin{aligned} W_0 x_0 &\geq h_0 \\ A_1 x_0 + W_1 x_1 &\geq h_1(\xi_1) \\ A_2 x_1 + W_2 x_2 &\geq h_2(\xi_2) \\ &\vdots \\ A_T x_{T-1} + W_T x_T &\geq h_T(\xi_T) \qquad (1) \\ x_1 &\lhd \mathcal{F}_1 \\ &\vdots \\ x_T &\lhd \mathcal{F}_T. \end{aligned}$$

Nomenclature

c	costs
W	recourse matrices
A	technology matrices
h	right hand sides

We distinguish between:

- random costs
- random recourse matrices
- random technology matrices
- random right hand sides

and all combinations.

Properties of utility functionals

Y is a profit variable.
$Y \mapsto \mathcal{U}(Y)$ is called an *utility functional* (negative risk functional) if it satisfies the following conditions for all profit variables Y:

- $\mathcal{U}(Y + c) = \mathcal{U}(Y) + c$ (*translation-equivariance*, cash-invariance)
- $\mathcal{U}(\lambda Y + (1-\lambda)\tilde{Y}) \geq \lambda \mathcal{U}(Y) + (1-\lambda)\mathcal{U}(\tilde{Y})$ (*concavity*),
- $Y \leq \tilde{Y}$ implies $\mathcal{U}(Y) \leq \mathcal{U}(\tilde{Y})$ (*monotonicity*).

Properties of risk functionals

Y is again a profit variable, if the problem is written for a loss/cost variable L, set $Y = -L$.

- $\mathcal{R}(Y + c) = \mathcal{R}(Y) - c$ (*translation-antivariance*, cash-antivariance)
- $\mathcal{R}(\lambda Y + (1-\lambda)\tilde{Y}) \leq \lambda \mathcal{R}(Y) + (1-\lambda)\mathcal{R}(\tilde{Y})$ (*convexity*),
- $Y \leq \tilde{Y}$ implies $\mathcal{R}(Y) \geq \mathcal{R}(\tilde{Y})$ (*monotonicity*).

More Definitions

An utility/acceptability functional $\mathcal{U}$ is called

- *positively homogeneous* if

$$\mathcal{U}(\lambda Y) = \lambda \mathcal{U}(Y), \quad \forall \lambda \geq 0$$

- *version-independent* (law-invariant) if $\mathcal{U}(Y)$ depends only on the distribution function $G_Y(u) = \mathbb{P}\{Y \leq u\}$ of Y.

If $\mathcal{U}$ is version independent, then the monotonicity property can be written as

$$Y^{(1)} \prec_{FSD} Y^{(2)} \text{ implies that } \mathcal{U}(Y^{(1)}) \leq \mathcal{U}(Y^{(2)})$$

Here $\prec_{FSD}$ means *first order stochastic dominance*. In some cases the functional $\mathcal{U}$ is even monotonic w.r.t.*second order dominance*

$$Y^{(1)} \prec_{SSD} Y^{(2)} \text{ implies that } \mathcal{U}(Y^{(1)}) \leq \mathcal{U}(Y^{(2)})$$

Order relations

Definition: Orderings (Fishburn (1980).
Let $Y^{(1)}, Y^{(2)}$ be profit&loss variables, not necessarily defined on the same probability space.

(i) $Y^{(2)}$ dominates $Y^{(1)}$ in the first order sense (in symbol $Y^{(1)} \prec_{FSD} Y^{(2)}$, if

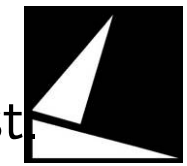
$$\mathbb{E}[U(Y^{(1)})] \leq \mathbb{E}[U(Y^{(2)})]$$

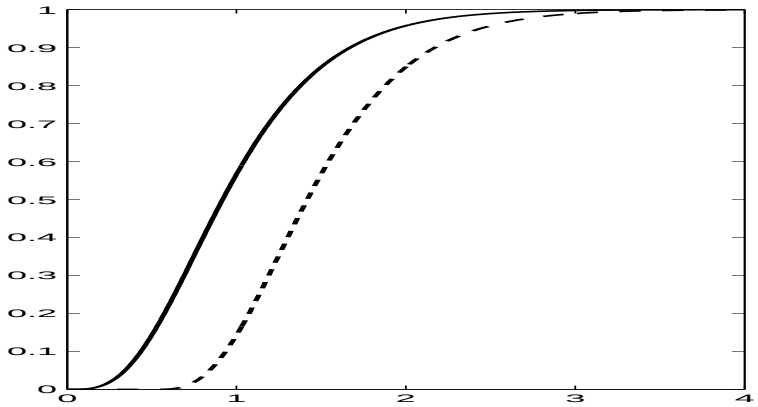
for all nondecreasing utility functions U, for which both integrals exist.

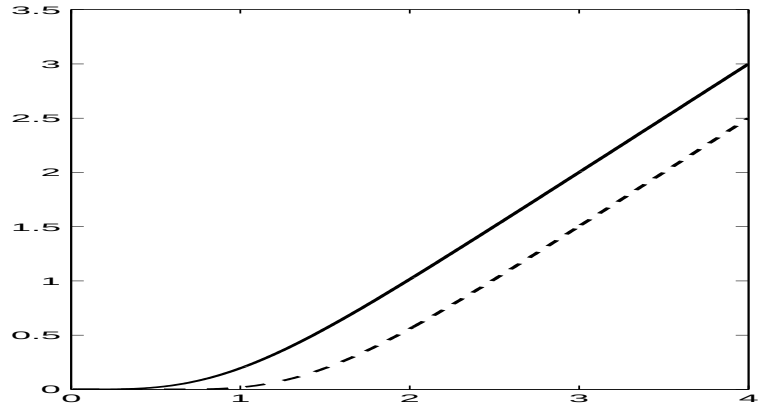
(ii) $Y^{(2)}$ dominates $Y^{(1)}$ in the second order sense (in symbol $Y^{(1)} \prec_{SSD} Y^{(2)}$, if

$$\mathbb{E}[U(Y^{(1)})] \leq \mathbb{E}[U(Y^{(2)})]$$

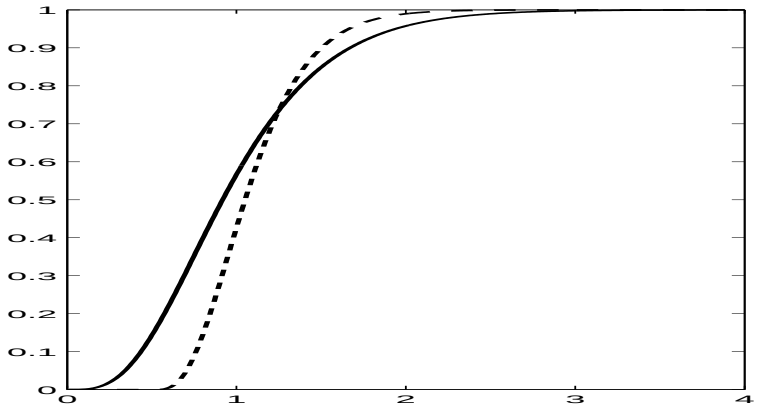
for all nondecreasing concave U, for which both integrals exist.

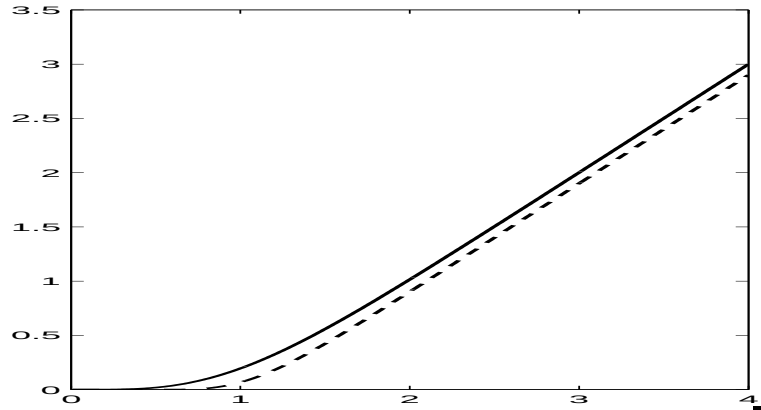




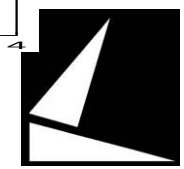


Left: the distribution functions G_1 of $Y^{(1)}$ (solid) and G_2 of $Y^{(2)}$ (dashed) Right:the integrated distribution functions $\mathcal{G}_1(x) = \int_{\infty}^{x} G_1(t)\,dt$ of $Y^{(1)}$ (solid) and $\mathcal{G}_2$ of $Y^{(2)}$ (dashed); The relation $Y^{(1)} \prec_{FSD} Y^{(2)}$ holds.





$Y^{(1)} \prec_{SSD} Y^{(3)}$ holds, but $Y^{(1)} \prec_{FSD} Y^{(3)}$ does not hold.



"Coherent risk measures"

Given an utility functional $\mathcal{U}$, the mappings

$$\mathcal{R} := -\mathcal{U}$$

are called the associated *risk functional/ risk measure*.
Coherent risk functionals are negative utility functionals which are positively homogeneous in addition. (Artzner et al., 1999).

Utility type functionals

Let U be a concave, strictly monotonic utility function.

$$\mathcal{U}(Y) = U^{-1}(\mathbb{E}[U(Y)]).$$

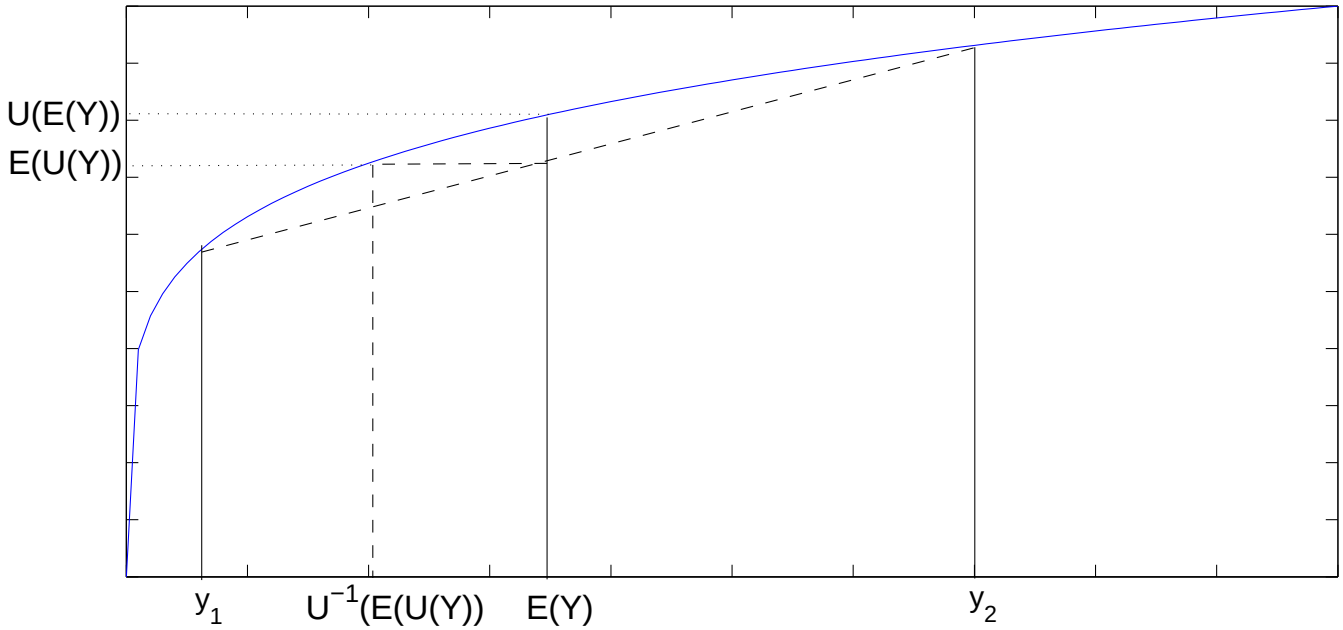
Then $\mathcal{U}$ is the *certainty equivalent* and

$$\mathcal{D}(Y) = \mathbb{E}Y - \mathcal{U}(Y),$$

may be seen as *risk premium*. The decision maker with utility function U is indifferent between Y and the deterministic value $\mathcal{U}$ in the sense that

$$\mathbb{E}[U(Y)] = U[\mathcal{U}(Y)].$$

This type of functionals is translation-equivariant iff $U(x) = -k\exp(-\gamma x) + d$; $k \geq 0$ or $U(x) = kx + d$. Up to affine transformations, these are the entropic functionals and the expectation itself.


U(E(Y))
E(U(Y))
y1
U−1(E(U(Y))
E(Y)
y2

Risk premium in insurance

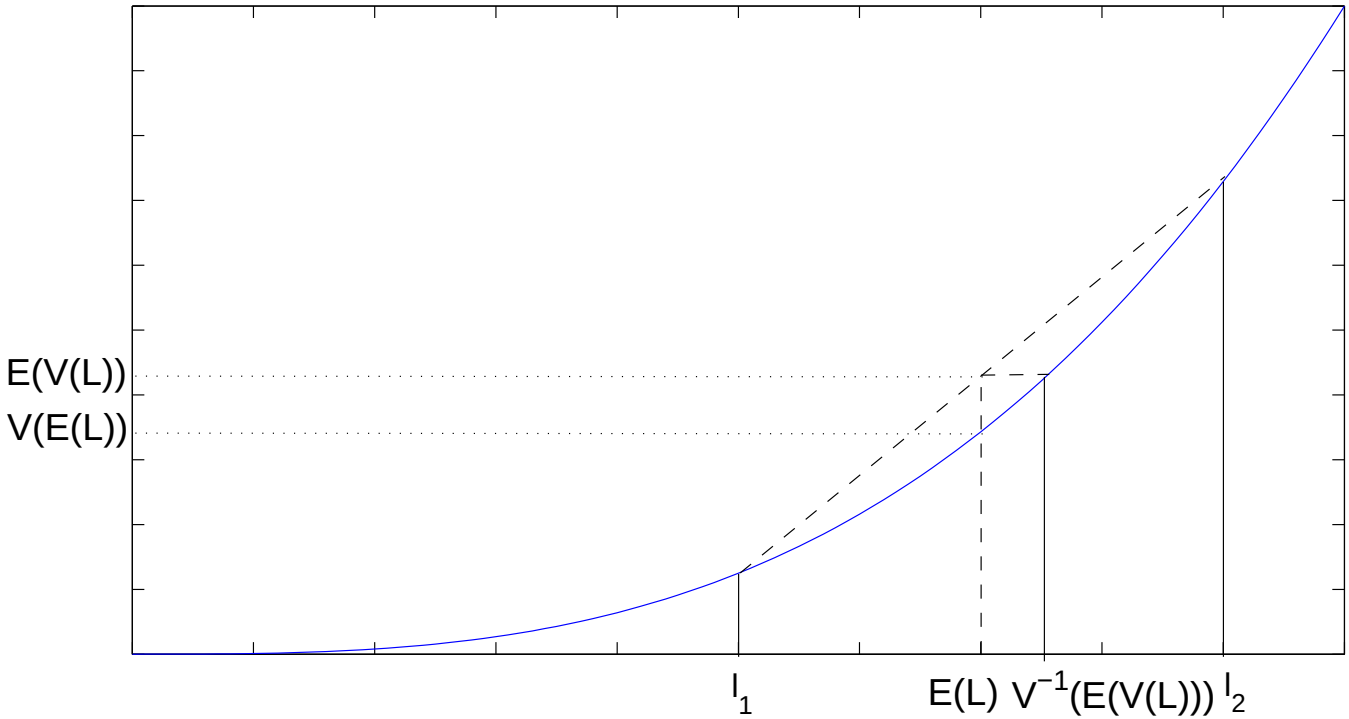
If L is a loss variable and let V be a convex, strictly monotonic disutility function. Then the CE premium π is calculated as

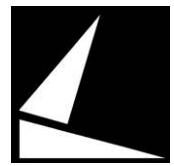
$$\pi(L) = V^{-1}(\mathbb{E}[V(L)]) \geq \mathbb{E}(L)$$

and

$$\pi(L) - \mathbb{E}(L)$$

is the *risk premium*. Examples for risk premia are $V(u) = u^q$ for $q > 1$ and $V(u) = \exp(\gamma u)$ for $\gamma > 0$.


E(V(L))
V(E(L))
l_1
E(L)
$V^{-1}(E(V(L)))$
l_2



Risk aversion

The basic relation between the certainty equivalent and the risk aversion is given by

$$\begin{aligned} U^{-1}\mathbb{E}[U(Y)] &= \mathbb{E}(Y) + \frac{U''}{U'(\mathbb{E}(Y))} \cdot \frac{\mathbb{V}\mathrm{ar}(Y)}{2} + \text{ higher order terms} \\ &\sim \mathbb{E}(Y) - ARA(\mathbb{E}(Y)) \cdot \frac{\mathbb{V}\mathrm{ar}(Y)}{2} \end{aligned}$$

with *absolute risk aversion (ARA)*

$$ARA(y) = -\frac{U''(y)}{U'(y)}$$

and *relative risk aversion (RRA)*

$$RRA(y) = -\frac{yU''(y)}{U'(y)}$$

Power, log- and exponential utility

$U(y)$	$ARA(y)$		$RRA(y)$	
$\frac{y^{1-\gamma}-1}{1-\gamma}$	$\frac{1}{y\gamma}$	DARA	$\frac{1}{\gamma}$	CRRA
$\log(y)$	$\frac{1}{y}$	DARA	1	CRRA
$-\exp(-y\gamma)$	γ	CARA	$y\gamma$	IRRA

The concave dual and the certainty equivalent:

$U(y)$	$U^+(z)$	$U^{-1}\mathbb{E}[U(Y)]$
$\frac{y^{1-\gamma}-1}{1-\gamma}$	$\frac{-\gamma}{1-\gamma}z^{1-1/\gamma}+\frac{1}{1-\gamma}$	$[\mathbb{E}(Y^{1-\gamma})]^{1/(1-\gamma)}$
$\log(y)$	$1+\log(z)$	$\exp(\mathbb{E}[\log(Y)])$
$-\exp(-y\gamma)$	$\frac{z}{\gamma}(1-\log(z/\gamma)$	$-\frac{1}{\gamma}\log\mathbb{E}[\exp(-\gamma Y)]$

By the Rockafellar-Fenchel-Moreau Theorem, every concave upper semicontinuous (u.s.c.) functional $\mathcal{U}$ has a representation of the form

$$\mathcal{U}(Y) = \inf_Z \{\mathbb{E}(Y\,Z) - \mathcal{U}^+(Z)\}, \tag{2}$$

where $\mathcal{U}^+(Z) = \inf_Y \{\mathbb{E}(Y\,Z) - \mathcal{U}(Y)\}$ is the *conjugate* of $\mathcal{U}$. We call (2) a *dual representation* and $dom(\mathcal{U}^+) = \{Z : \mathcal{U}^+(Z) > -\infty\}$ the *set of supergradients*. Notice that if Z is a supergradient,

$$\mathcal{U}(Y) \leq \mathbb{E}(Y\,Z) - \mathcal{U}^+(Z),$$

i.e. the affine-linear functional $Y \mapsto \mathbb{E}(Y\,Z) - \mathcal{U}^+(Z)$ is a majorant of $\mathcal{U}$. The Fenchel-Moreau inequality

$$\mathbb{E}(Y\,Z) \geq \mathcal{U}(Y) + \mathcal{U}^+(Z) \tag{3}$$

follows.

Example: The entropic functional

- *Primal form*

$$\mathcal{U}(Y) = -\frac{1}{\gamma}\log \mathbb{E}[\exp(-\gamma Y)].$$

- *Dual form*

$$\mathcal{U}(Y) = \inf\{\mathbb{E}(Y\,Z) + \frac{1}{\gamma}\mathbb{E}(Z\log Z) : \mathbb{E}(Z) = 1, Z \geq 0\}.$$

The entropic functional is monotonic w.r.t. $\prec_{SSD}$. It is the Certainty equivalent for the exponential utility.

Example: The average value-at-risk

- *Primal form.* $\mathbb{AV@R}_\alpha(Y) = \frac{1}{\alpha}\int_0^\alpha G_Y^{-1}(p)\,dp$

$$\mathbb{AV@R}_0(Y) = \operatorname{ess-inf}(Y).$$

- *Dual form*

$$\mathbb{AV@R}_\alpha(Y) = \inf\{\mathbb{E}(Y\,Z) : \mathbb{E}(Z) = 1, 0 \leq Z \leq 1/\alpha\}.$$

Other names for this functional: *conditional value-at-risk* (Rockefellar and Uryasev (2002)), *expected shortfall* (Acerbi and Tasche (2002)) and *tail value-at-risk* (Artzner et al. (1999)). The name average value-at-risk is due to Föllmer and Schied (2004). The $\mathbb{AV@R}$ is monotonic w.r.t. $\prec_{SSD}$.

Risk corrected expectation I

Let h be a nonnegative convex function on $\mathbb{R}$ with $h(0) = 0$. .

- *Primal form.* $\mathcal{U}(Y) = \mathbb{E}Y - \mathbb{E}[h(Y - \mathbb{E}Y)]$.
- *Dual form.* $\mathcal{U}(Y) = \inf\{\mathbb{E}(Y\,Z) + D_{h^*}(Z) : \mathbb{E}Z = 1\}$, where $D_{h^*}(Z) = \inf\{\mathbb{E}[h^*(Z - a)] : a \in \mathbb{R}\}$and $h^*(u) = \sup\{uv - h(v) : v \in \mathbb{R}\}$ is the Fenchel conjugate of h.

For example $h(u) = u^2$. However, for every $\delta > 0$, there are random variables $Y^{(1)}$ and $Y^{(2)}$ such that $Y^{(1)} \prec_{FSD} Y^{(2)}$, but $\mathbb{E}Y^{(1)} - \delta\mathbb{V}\mathrm{ar}Y^{(1)} > \mathbb{E}Y^{(2)} - \delta\mathbb{V}\mathrm{ar}Y^{(2)}$.

Risk corrected expectation II

- *Primal form*

$$\mathcal{U}(Y) = \mathbb{E}Y - \inf\{\mathbb{E}[h(Y-a)] : a \in \mathbb{R}\}.$$

- *Dual form*

$$\mathcal{U}(Y) = \inf\{\mathbb{E}(Y\,Z) + \mathbb{E}[h^*(1-Z)] : \mathbb{E}(Z) = 1\}$$

where $D_{h^*}(Z) = \inf\{\mathbb{E}[h^*(Z-a)] : a \in \mathbb{R}\}$.

Distortion functionals

Distortion functionals were introduced independently as insurance pricing principles (Deneberg (1989), Wang (2000)) and by Yaari (1987) (Yaari's dual functionals).

- *Primal form*

$$\mathcal{U}(Y) = \int_0^1 G_Y^{-1}(p)\, h(p)\, dp$$

where G_Y is the distribution function of Y.

- *Dual form*

$$\mathcal{U}(Y) = \inf\{\mathbb{E}(Y\, Z) : \mathbb{E}(\phi(Z)) \leq \int \phi(h(u))\, du, \phi \text{ convex }, \phi(0) = 0\}.$$

Multistage stochastic programs

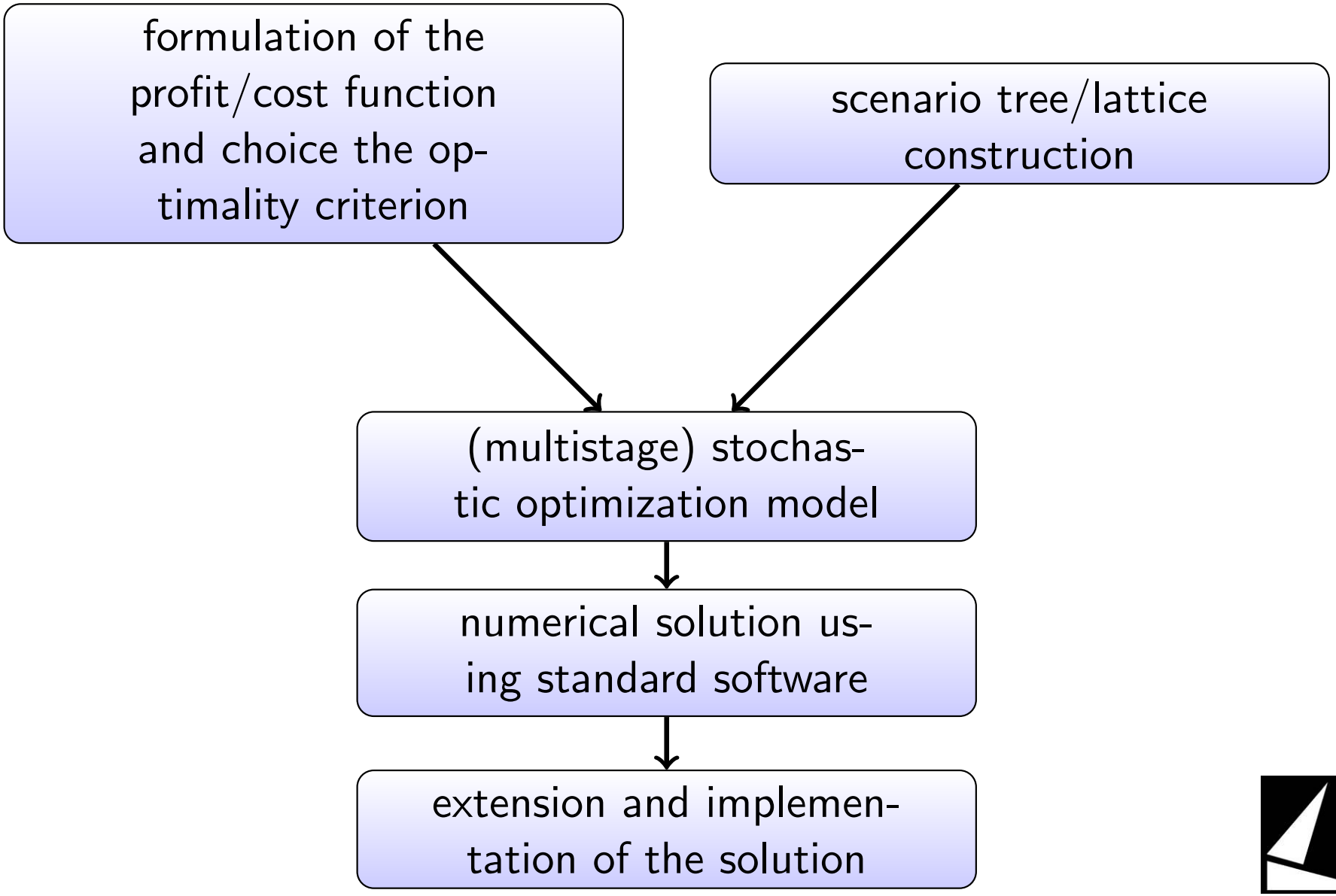
$$\min\{\mathcal{R}[Q(x_0, \xi_1, \ldots, x_{T-1}, \xi_T)] : x \lhd \mathcal{F} x \in \mathbb{X}\},$$

where

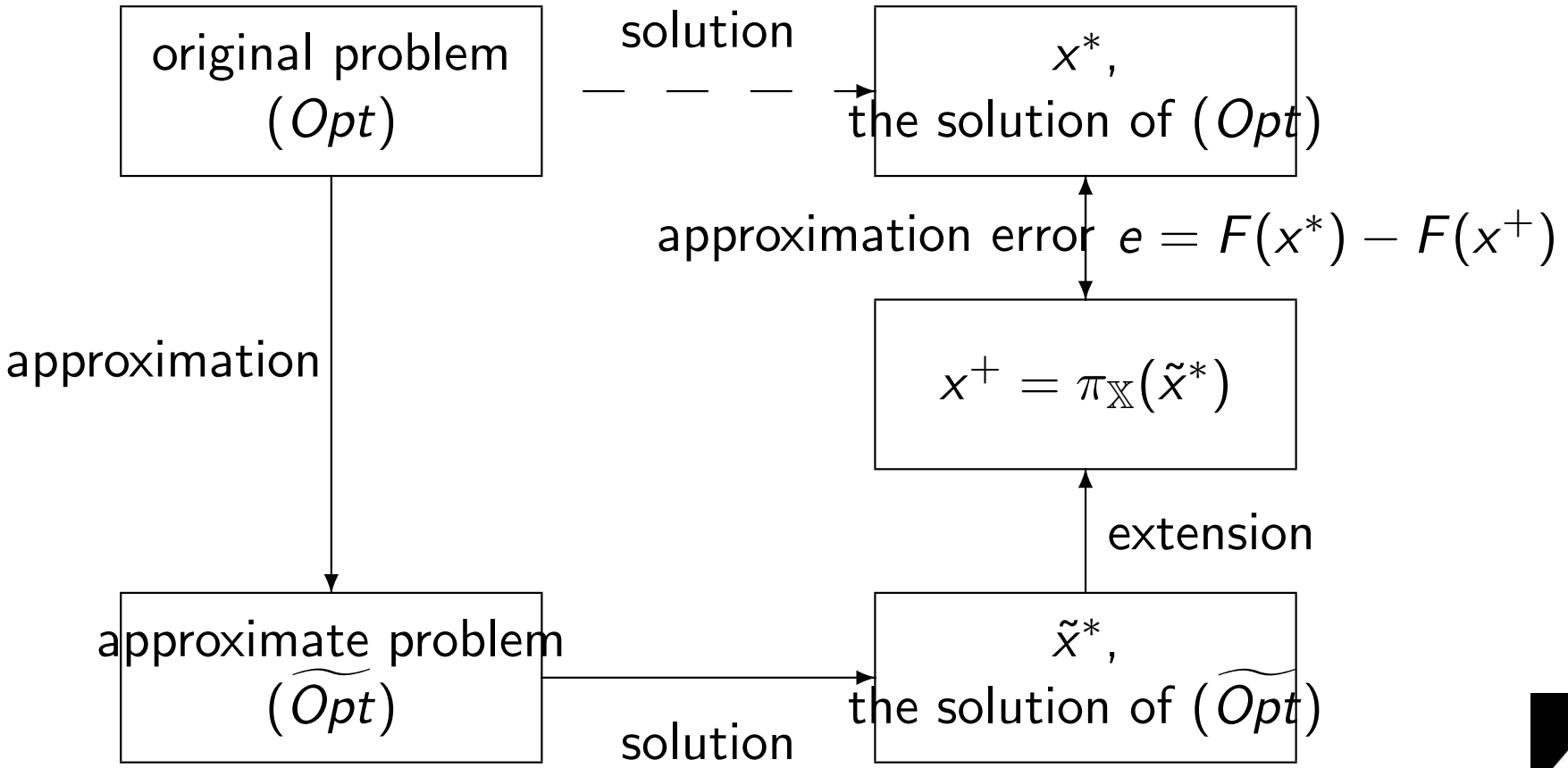
$\xi = (\xi_1, \ldots, \xi_T)$	a random scenario process defined on $(\Omega, \mathcal{F}, P)$
$x = (x_0, \ldots, x_{T-1})$	the sequence of decisions,
$Q(x_0, \xi_1, \ldots, \xi_T)$	the profit function,
$\mathcal{F} = (\mathcal{F}_1, \ldots \mathcal{F}_T)$	a filtration (an increasing sequence of σ-algebras),
$\xi \lhd \mathcal{F}$	ξ is adapted to $\mathcal{F}$, i.e. $\sigma(\xi) \subseteq \mathcal{F}$
$x \lhd \mathcal{F}$	the nonanticipativity condition.

In general, a multistage stochastic program is a variational problem, since the solutions must be found among functions and not - as usual - among vectors. We have to approximate the problem by a simpler one.

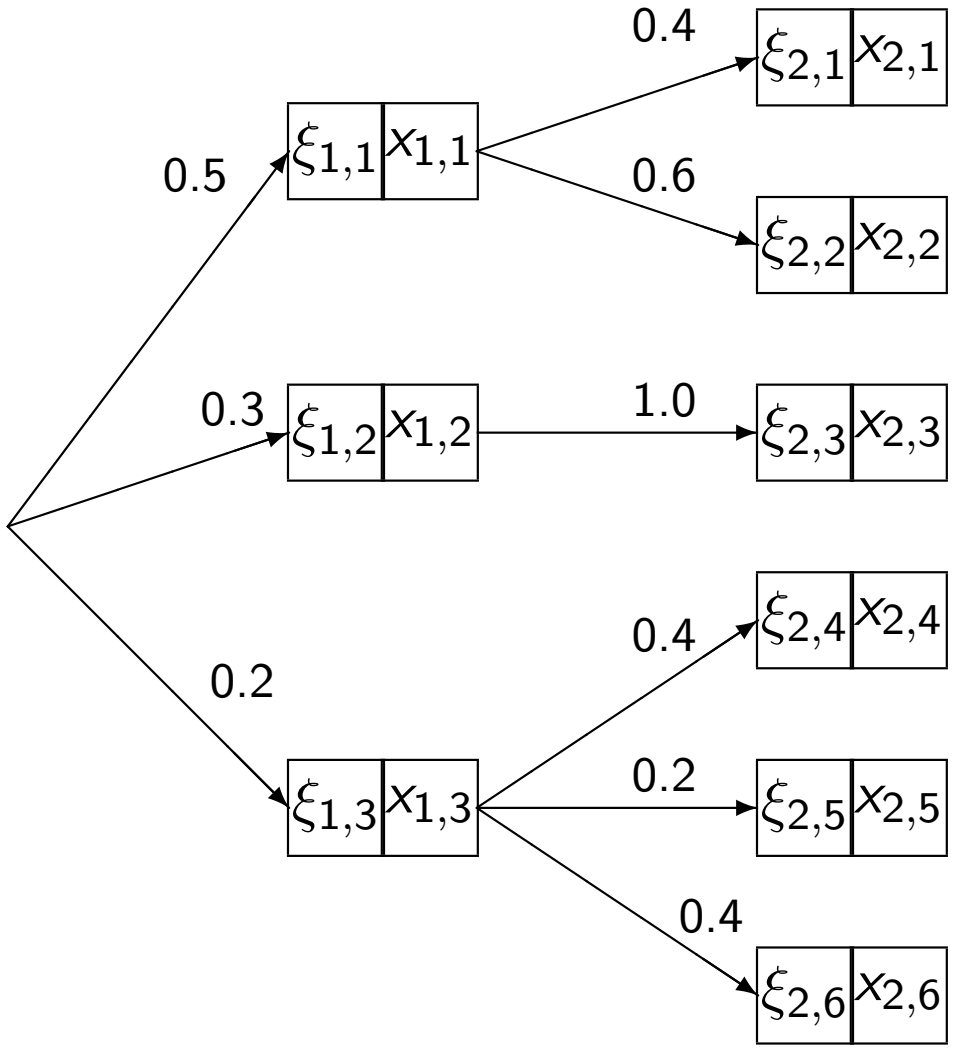
The phases of stochastic optimization

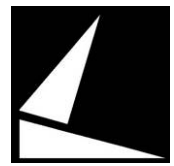


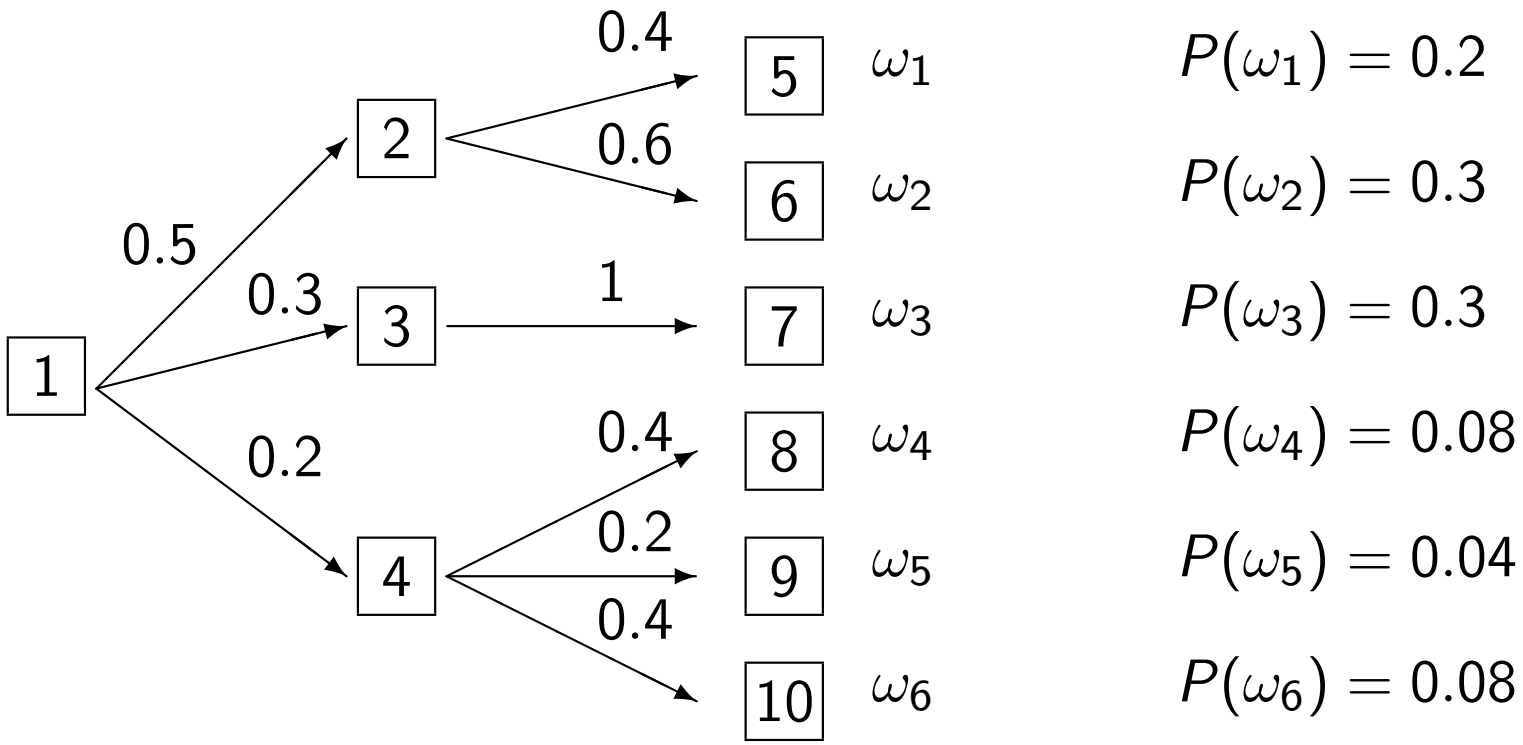
Approximation of stochastic decision processes



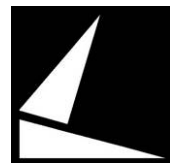
Trees encode finite valued stochastic processes and finite filtrations



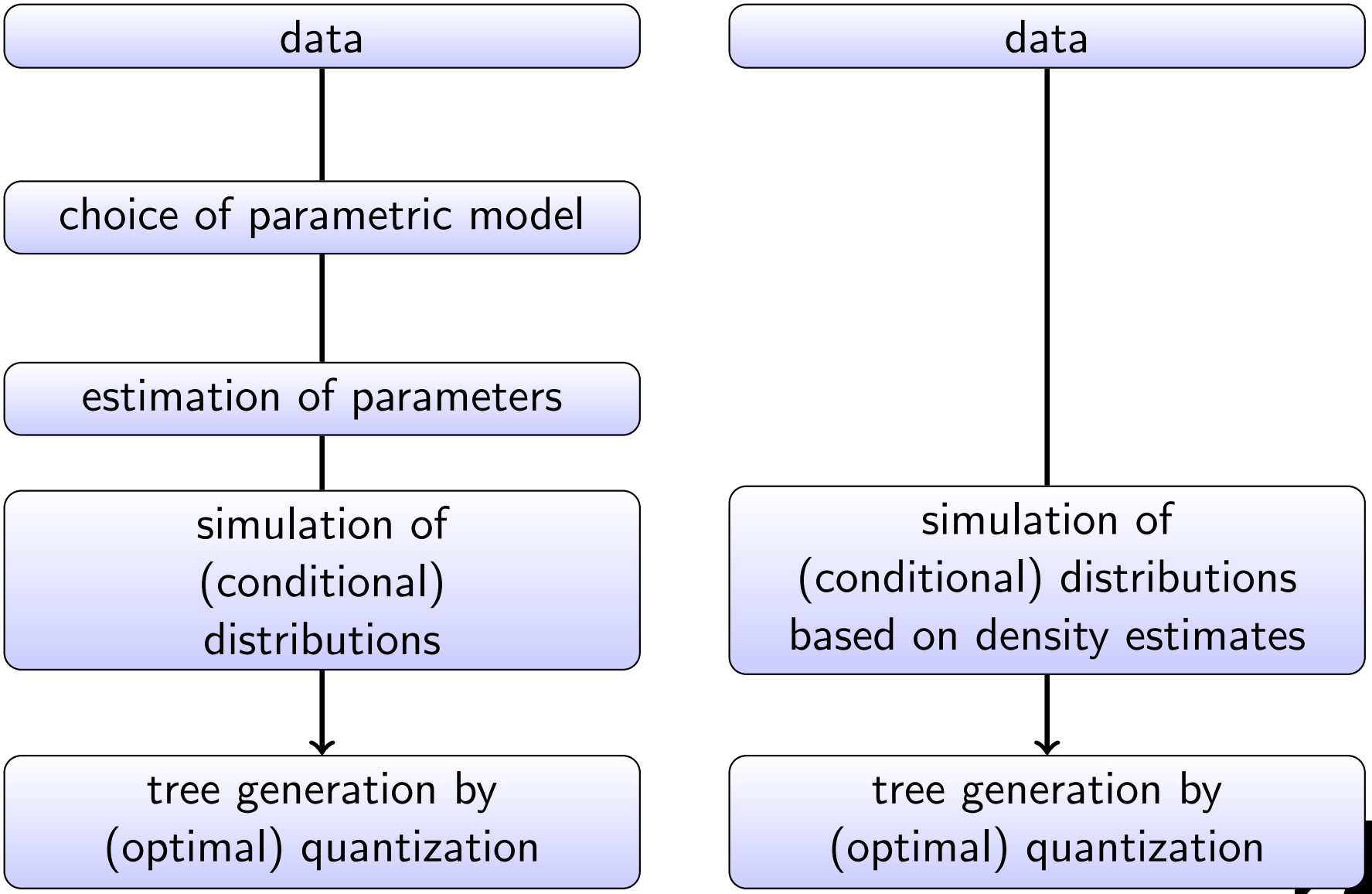




An exemplary finite tree process $\nu = (\nu_0, \nu_1, \nu_2)$ with nodes $\mathcal{N} = \{1, \ldots 10\}$ and leaves $\mathcal{N}_2 = \{5, \ldots 10\}$ at $T = 2$ stages. The filtrations, generated by the respective atoms, are
$\mathcal{F}_2 = \sigma\left(\{\omega_1\}, \{\omega_2\}, \ldots \{\omega_6\}\right)$,
$\mathcal{F}_1 = \sigma\left(\{\omega_1, \omega_2\}, \{\omega_3\}, \{\omega_4, \omega_5, \omega_6\}\right)$ and
$\mathcal{F}_0 = \sigma\left(\{\omega_1, \omega_2, \ldots \omega_6\}\right)$



Scenario tree generation



The parametric (left) and the nonparametric approach (right) for scenario tree generation

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.

The Monge/Kantorovich/Wasserstein transportation istance

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

Theorem (Kantorovich-Rubinstein). Dualization:

$$d_1(P_1, P_2) = \inf\{\mathbb{E}(\|X - Y\| : (X, Y) \text{ is a bivariate r.v. with given marginal distributions } P_1 \text{ and } P_2\}.$$

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

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

The infimum is attained. The bivariate distribution with marginals P_1 resp. P_2 which is the minimizer is called the optimal

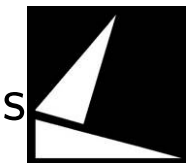
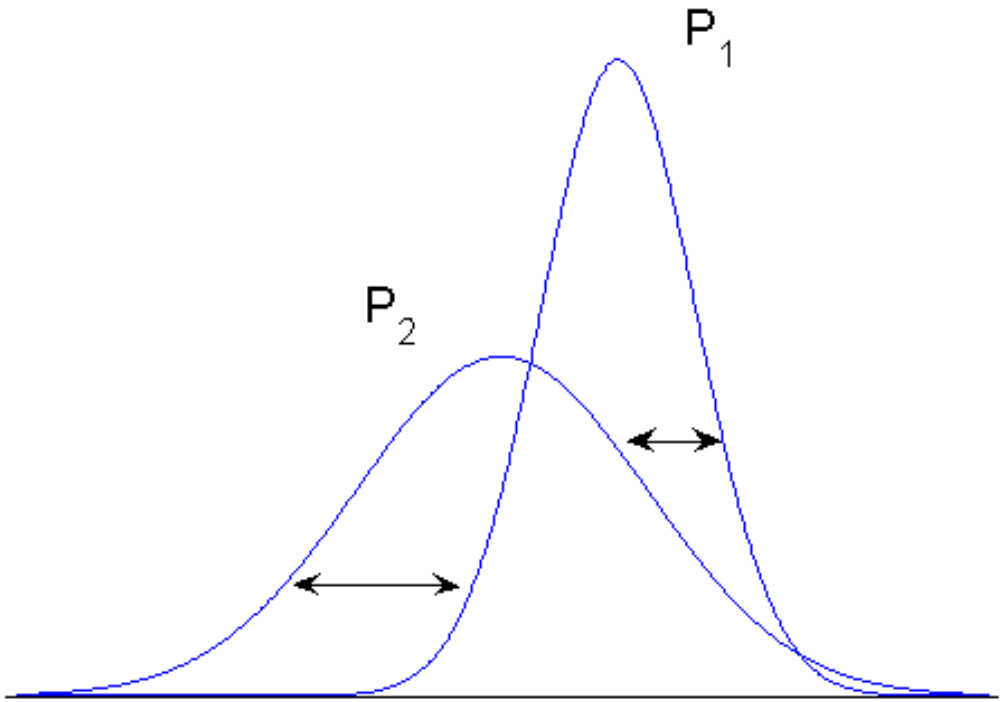


Illustration of the transportation distance



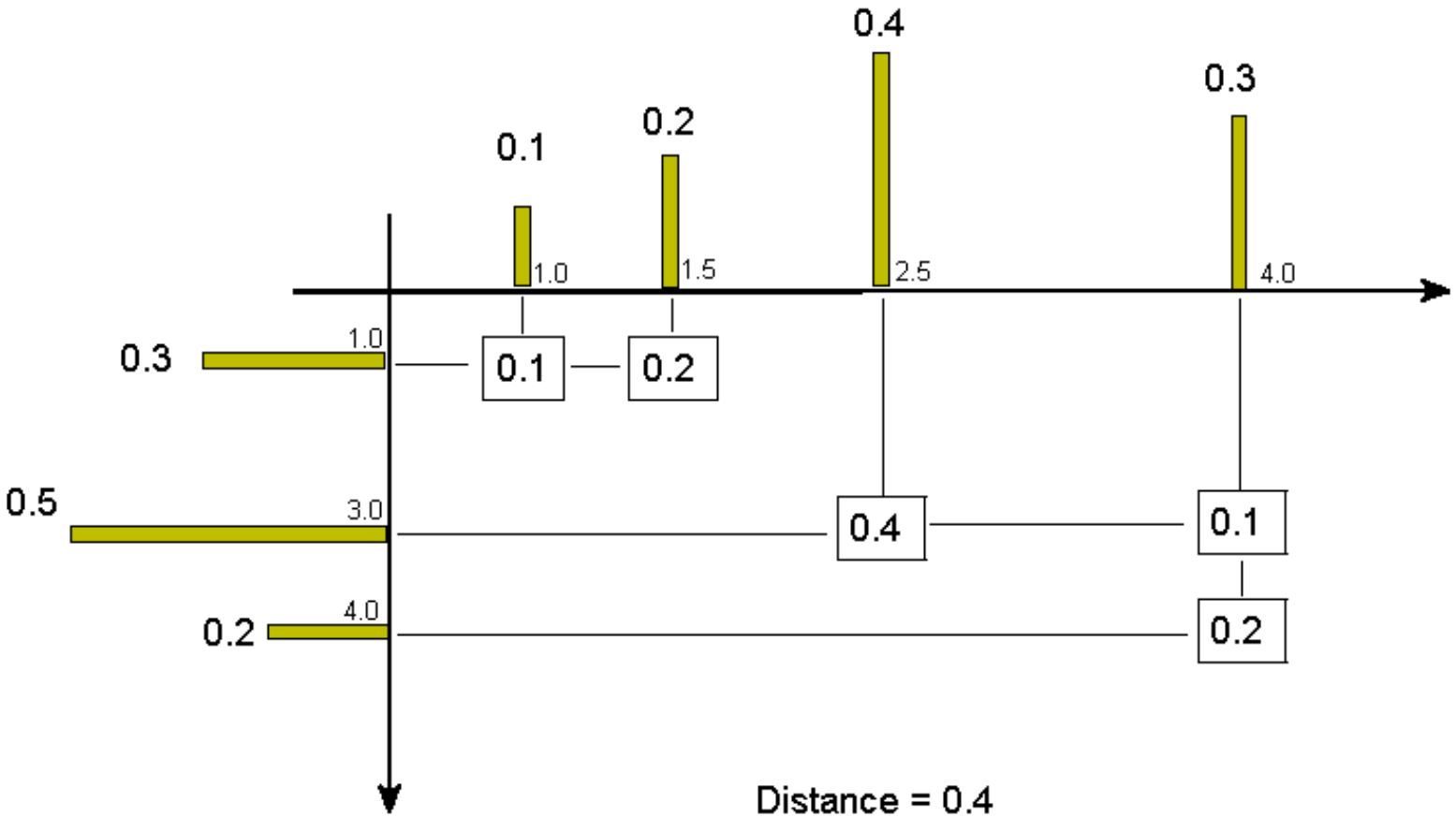
Remark. If both measures sit on a finite number of mass points $\{z_1, z_2, \ldots z_s\}$, then $d_r^r(P_1, P_2)$ is the optimal value of the following linear optimization problem:

$$\left\| \begin{array}{ll} \text{Maximize } \sum_i y_i(P_1(i) - P_2(i)) & \\ y_i - y_j \leq d(z_i, z_j) & \text{for all } i, j \end{array} \right.$$

or of its dual:

$$\left\| \begin{array}{ll} \text{Minimize } \sum_{i,j} \pi_{ij} d(z_i, z_j) & \\ \sum_i \pi_{ij} = P_1(j) & \text{for all } j \\ \sum_j \pi_{ij} = P_2(i) & \text{for all } i \end{array} \right.$$

Illustration: The Kantorovich distance as the solution of transportation problem



Historical 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).

Iterating: scenario trees are distributions of distributions

If (Ξ_1, d_1) and (Ξ_2, d_2) are metric 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 $\mathbb{R}^m$. Then we define the following spaces

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

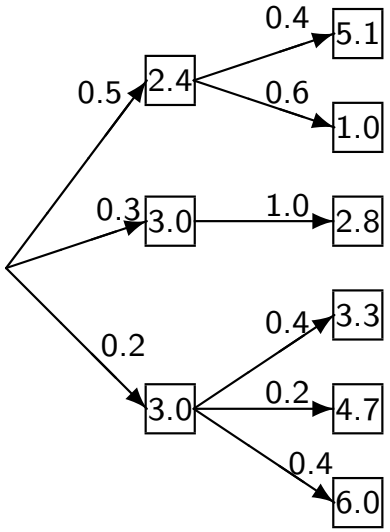
All spaces $\Xi_1, \ldots, \Xi_T$ are Polish spaces and they may carry probability distributions.

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:

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

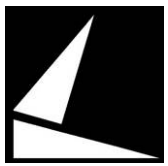
Examples for nested distributions



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

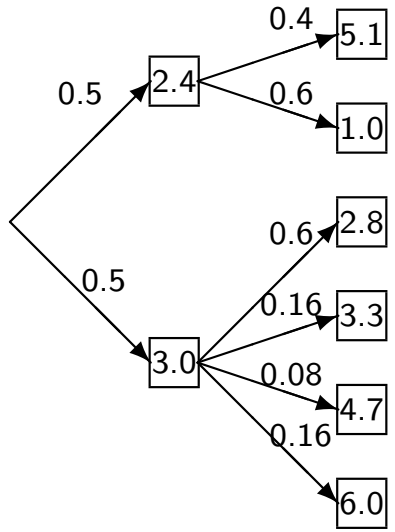
The embedded multivariate, but non-nested distribution of the scenario process can be gotten from it:

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



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 $\mathbf{d}(\tilde{\mathbb{P}}, \mathbb{P})$ is well defined. This distance makes sense, even if one process is discrete and the other is not.

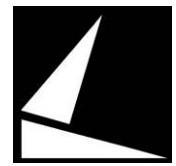
Theorem. Let $\mathbb{P}, \tilde{\mathbb{P}}$ be nested distributions and $P, \tilde{P}$ be the pertaining multiperiod distributions. Then

$$d(P, \tilde{P}) \leq \mathbf{d}(\mathbb{P}, \tilde{\mathbb{P}}).$$

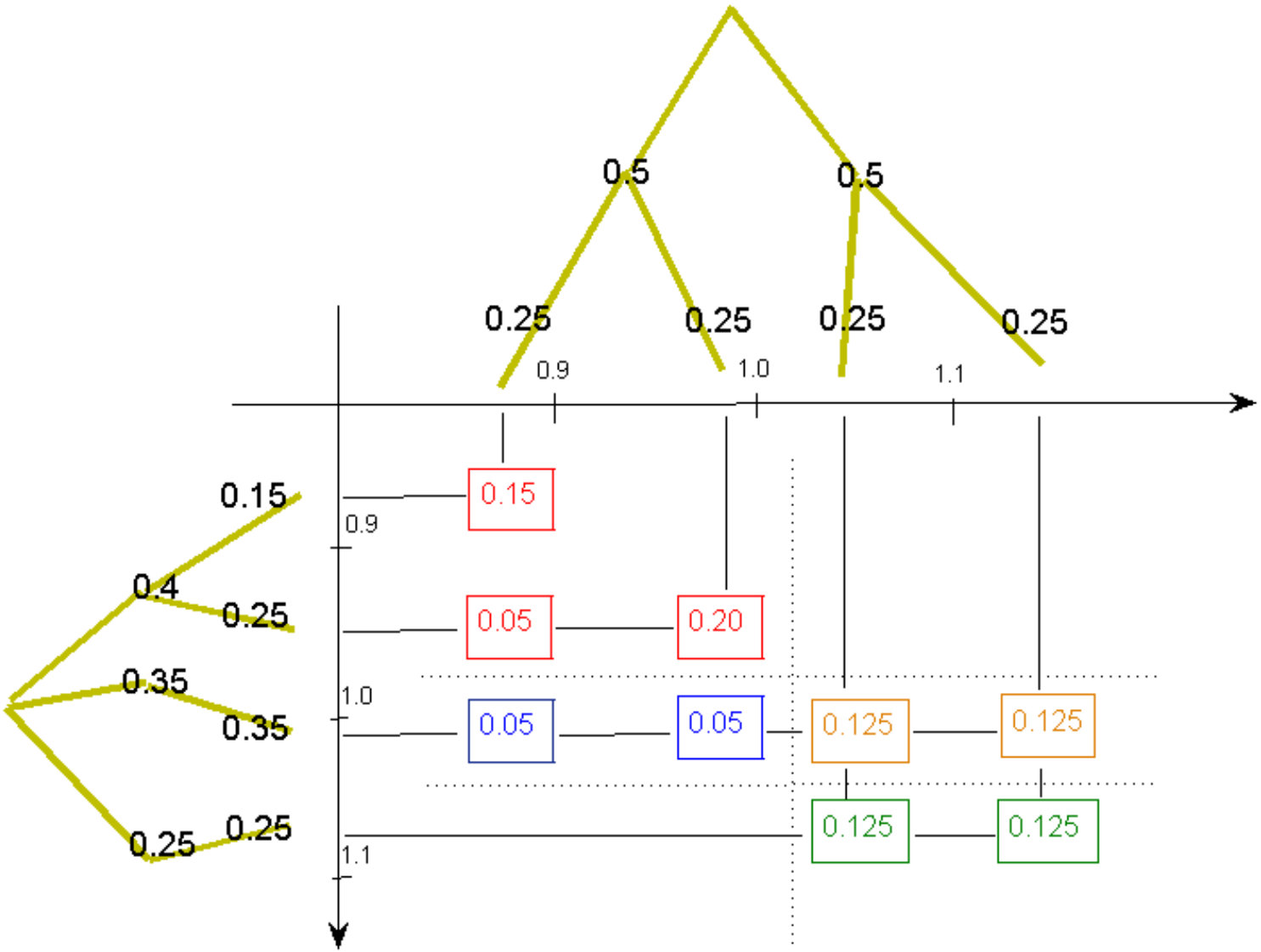
Theorem. For two nested distributions $\mathbb{P} := (\Xi, \mathcal{F}, P)$, $\tilde{\mathbb{P}} := \left(\tilde{\Xi}, \tilde{\mathcal{F}}, \tilde{P}\right)$ and a distance function on $d\colon \Xi \times \Xi' \to \mathbb{R}$ the *nested distance of order* $r \geq 1$ – denoted $\mathrm{dl}_r\left(\mathbb{P},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

$$\begin{array}{rll} \text{minimize} \\ (\text{in } \pi) & \left(\int d\left(\xi, \tilde{\xi}\right)^r \pi\left(\mathrm{d}\xi, \mathrm{d}\tilde{\xi}\right)\right)^{\frac{1}{r}} & \\ \text{subject to} & \pi\left(M \times \tilde{\Xi} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right) = P\left(M \mid \mathcal{F}_t\right) & \left(M \in \mathcal{F}_T\right) \\ & \pi\left(\Xi \times N \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t\right) = \tilde{P}\left(N \mid \tilde{\mathcal{F}}_t\right) & \left(N \in \tilde{\mathcal{F}}_T\right) \end{array} \tag{4}$$

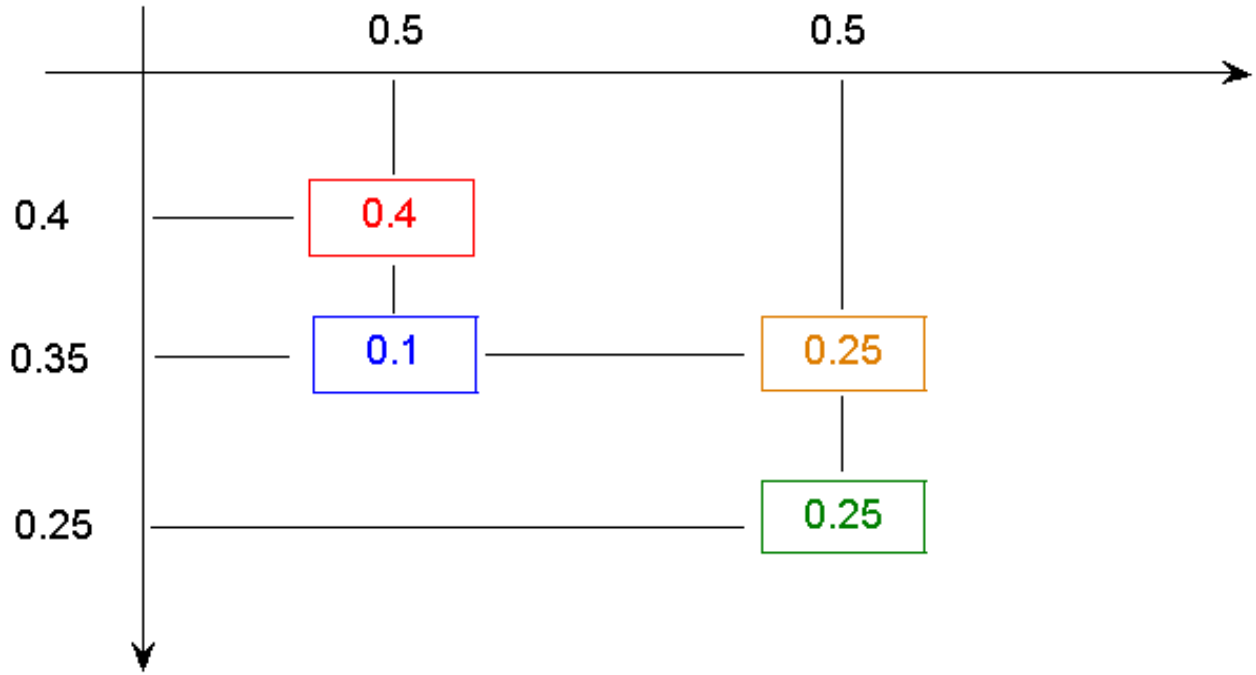
where the infimum in (4) is among all bivariate probability measures $\pi \in \mathcal{P}(\Xi \times \Xi')$, which are measures on the product sigma algebra $\mathcal{F}_T \otimes \tilde{\mathcal{F}}_T$. We will refer to the nested distance also as *process distance*, or *multistage distance*. The nested distance dl_2 (order $r = 2$), with d a weighted Euclidean distance is referred to as *quadratic nested distance*.



The nested distance: illustration



Distance = 0.758


0.5
0.5
0.4
0.4
0.35
0.1
0.25
0.25
0.25

What we have achieved

- The nested distribution contains both the information about the scenario values and the available information (the filtration) in a version-independent way.
- The nested distance quantifies the approximation error between continuous and discrete models (or large and small discrete models). It is the natural extension of the Kantorovich distance in two-stage models (see shortly).
- By minimaxing over nested balls, solutions which are robust against model ambiguity can be found (second part of the lecture).

How to calculate the nested distance

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

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

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

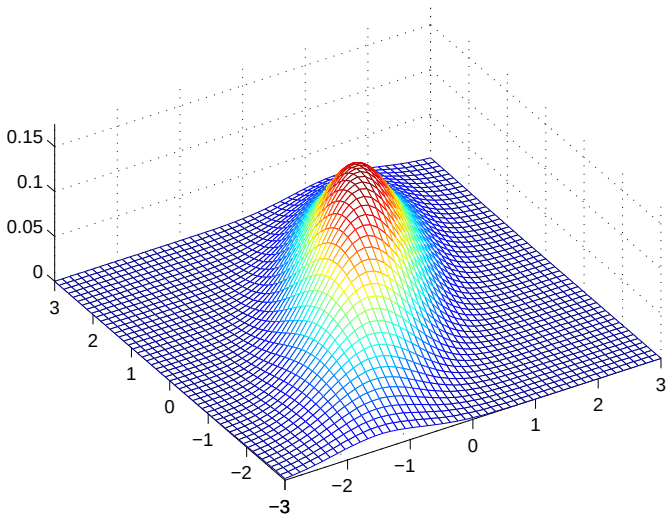
$$\mathbb{P} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right).$$

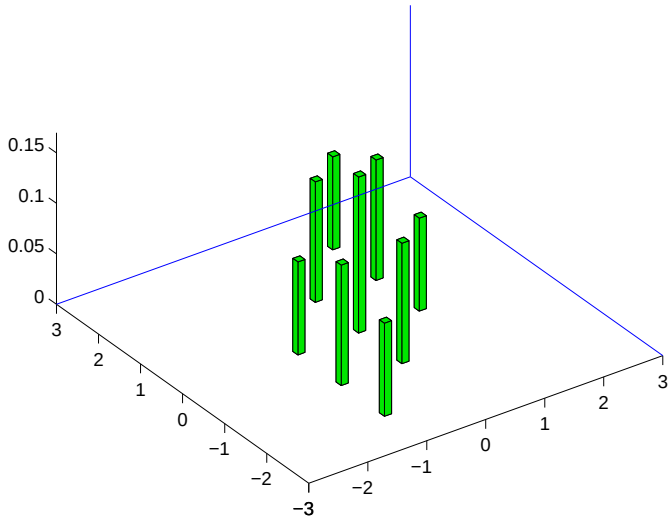
and

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

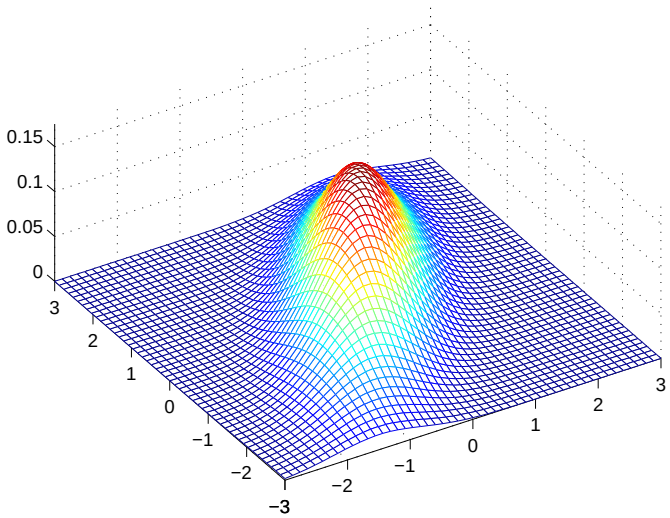
$$\left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline -2.058 & -1.029 & 0.0 \end{array}\right] & \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \hline -1.029 & 0.0 & 1.029 \end{array}\right] & \left[\begin{array}{ccc} 0.30345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

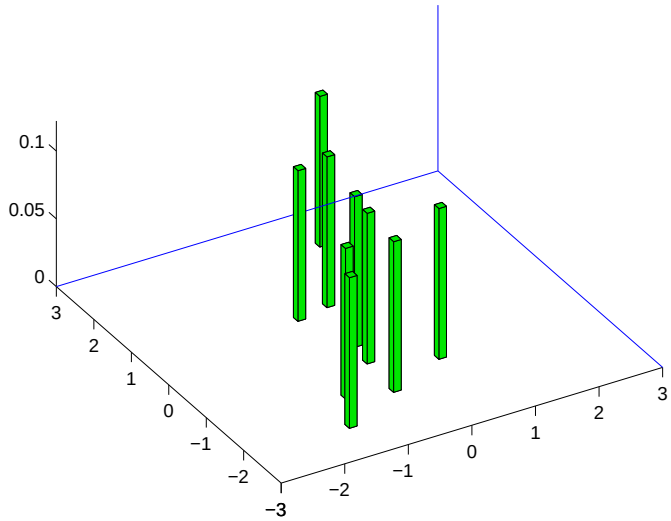
The nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 0.82$.
The distance of the multiperiod distributions is $d(P, \tilde{P}) = 0.68$.





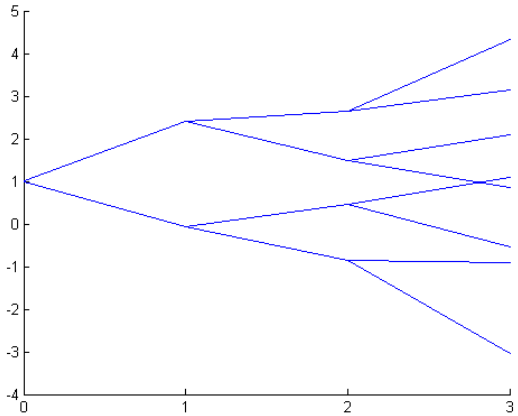
The nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 0.82$.
The distance of the multiperiod distributions is $d(P, \tilde{P}) = 0.68$.



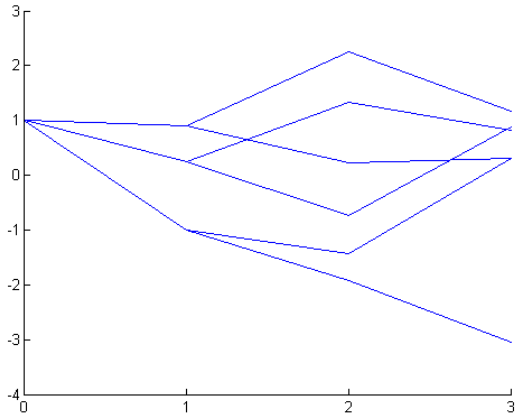


The nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 1.12$.
The distance of the multiperiod distributions is $d(P, \tilde{P}) = 0.67$.

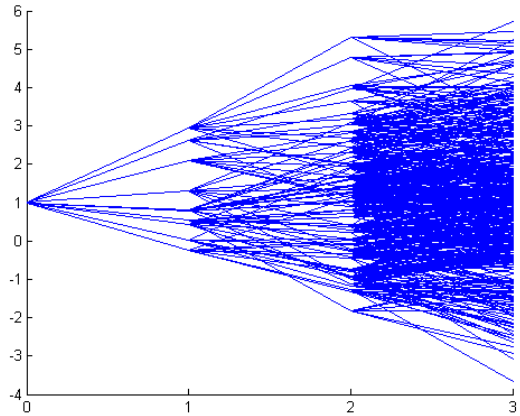
Examples of nested distances



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



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



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

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

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

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

The main approximation result

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

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

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

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

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

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

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

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

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

Remarks.

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

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

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

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

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

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

Optimal discretizations

Let P be a probability distribution on $\mathbb{R}^m$. The optimal discretization problem consists in finding s points $z_1, \ldots, z_s \in \mathbb{R}^m$ and probabilities $p_1, \ldots, p_s$ such that the discrete probability

$$\tilde{P} = \sum p_i \delta_{z_i}$$

minimizes

$$\mathrm{d}_r(P, \tilde{P})$$

among all probabilities sitting on at most m points.
Unfortunately, this is a nonconvex problem (see book by Graf and Luschgy). However, by global algorithms, one may solve the problem for some distributions, such as the multidimensional normal distribution (See the webpage of Gilles Pagès - the Pages-pages)

Scenario Tree Generation

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 two 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.

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

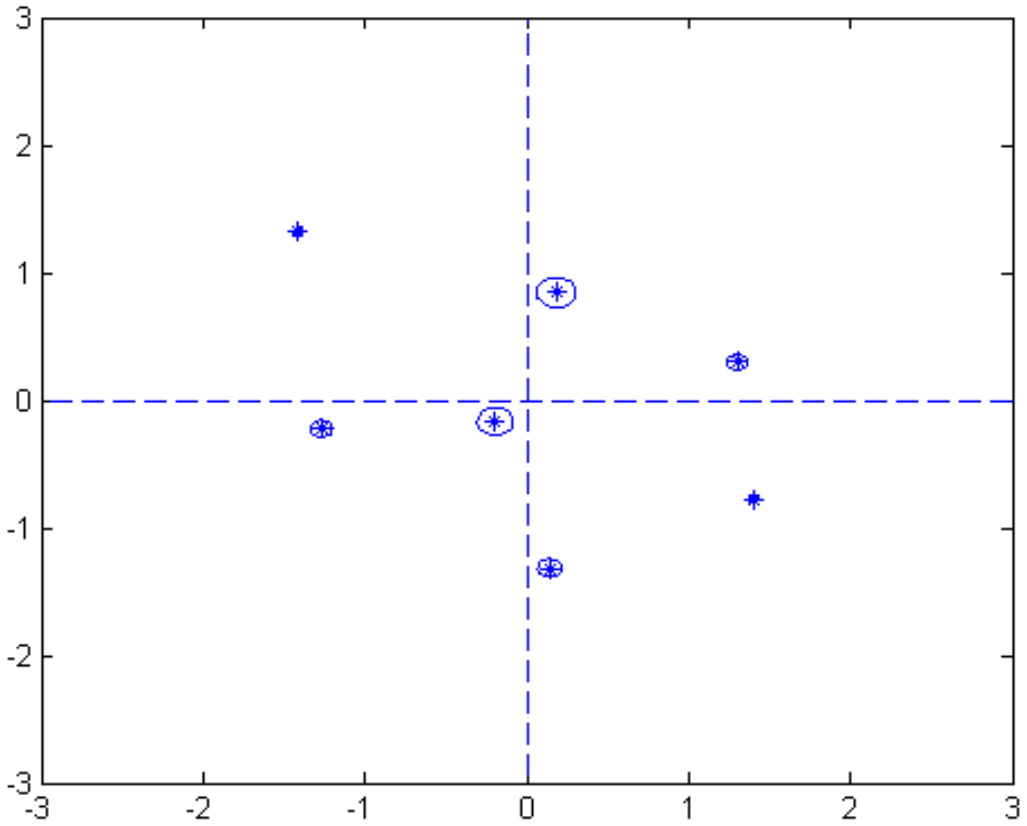
$$\begin{aligned} \tilde{\tilde{\Xi}}^{(0)} &= \{\tilde{\xi}_i^{(0)} : 1 \leq i \leq n\} \\ \tilde{p}_i^{(0)} &= 1/n \qquad \text{for } 1 \leq i \leq n \end{aligned}$$

2. Observe the next random value $\xi^{(k)}$
3. Find $j \in \{1, 2, \ldots, n\}$ such that $\xi^{(k)}$ is closest to $\tilde{\xi}_j^{(k)}$.
4. Set $\tilde{\xi}_j^{(k+1)} = \frac{k}{k+1}\tilde{\xi}_j^{(k)} + \frac{1}{k+1}\xi^{(k)}$ 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.

The tree reduction 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 $\tilde{\xi}_j^0$ of the approximating $\mathbb{P}'$-tree; $\mathbb{P}^0 := \tilde{\mathbb{P}}$.
- **Step 2 – Improve the quantizers**
  Find improved quantizers $\tilde{\xi}_j^{k+1}$:
  - 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**
  Setting $\pi \leftarrow \pi^k$ and $q \leftarrow q^{k+1}$ 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

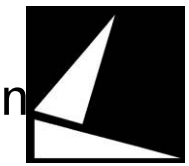
$$\mathrm{dl}_r^{k+1} < \mathrm{dl}_r^k - \varepsilon,$$

  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.

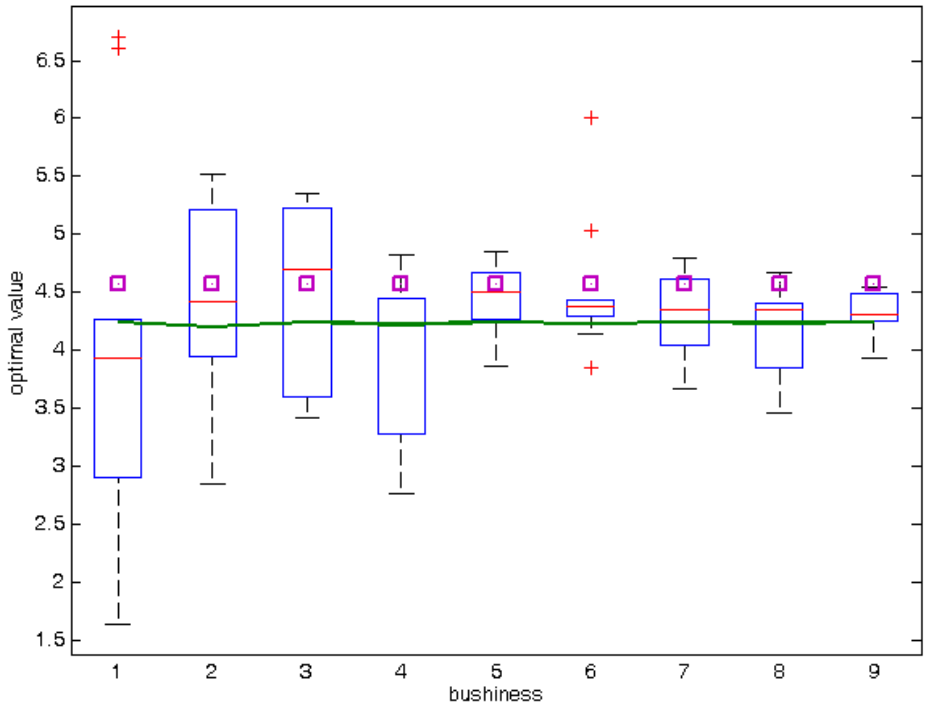
In case of the quadratic nested distance ($r = 2$) and the Euclidean distance the choice $\varepsilon = 0$ is possible.



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

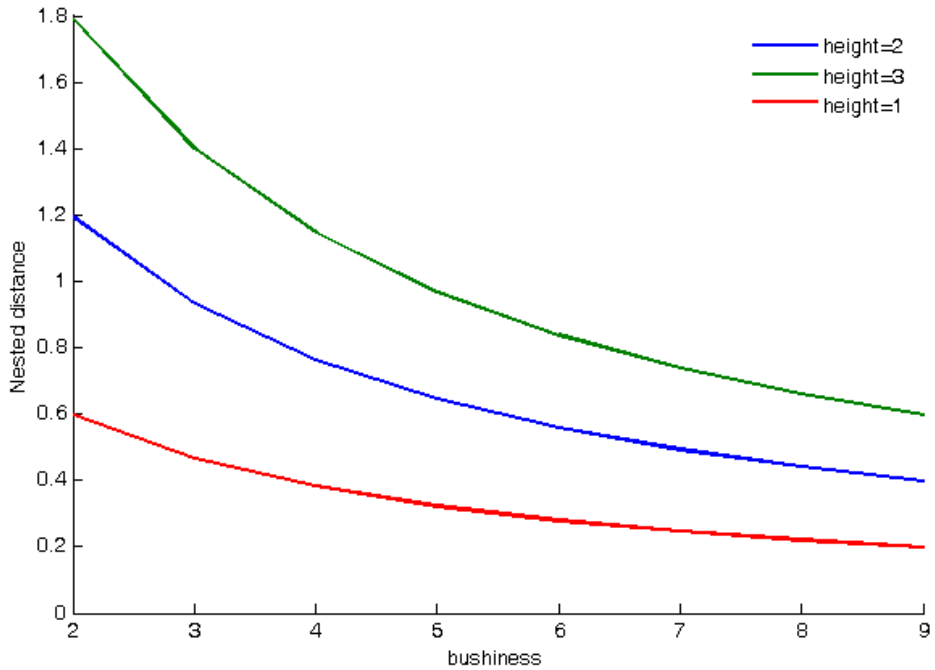
Monte Carlo sampling versus optimal quantification using nested distances



An inventory control problem (the multistage newsboy problem)

Approximation at work

Reducing the nested distance by making the tree bushier.



Time consistent decisions ?

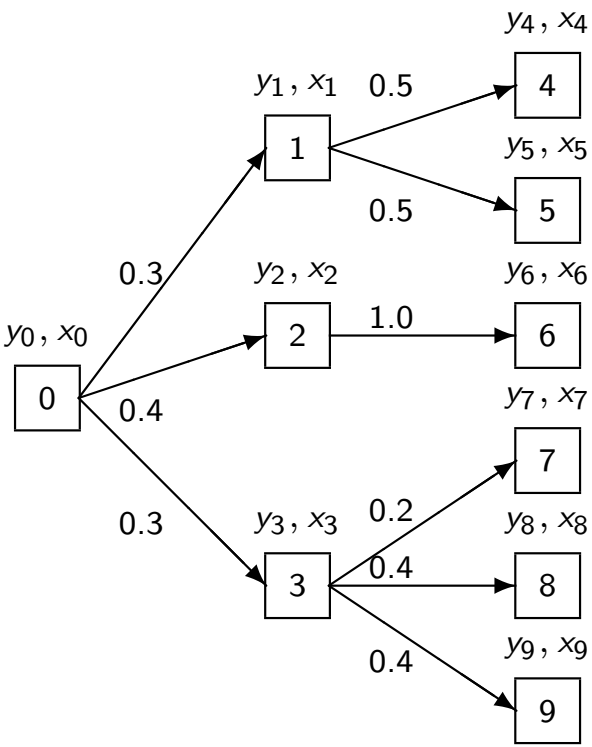
Let a stochastic multistage decision problem be given, which is defined on the basis of a tree process $\nu = (\nu_1, \ldots, \nu_T)$. Let $\mathbb{P}$ be the probability governing the tree process. Let $\mathbb{P}^{\nu_t = z}$ be the conditional distribution of the tree process, given that the value of ν_t is z. The solution is called time-consistent, if the solutions of the original problem and the conditional problems (when the decisions at times $1, \ldots, t-1$ are kept fixed) coincide on the subtree of $\nu_t = z$.

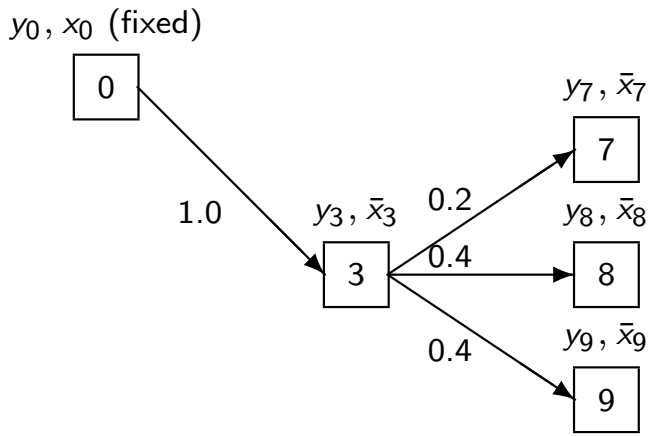
Proposition. If the objective is a nested acceptability functional (and no other constraints are present), then the decision problem leads to time consistent decisions.

y_i: values of the scenario process

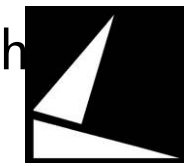
x_i: optimal decisions

$\boxed{i}$: node numbers



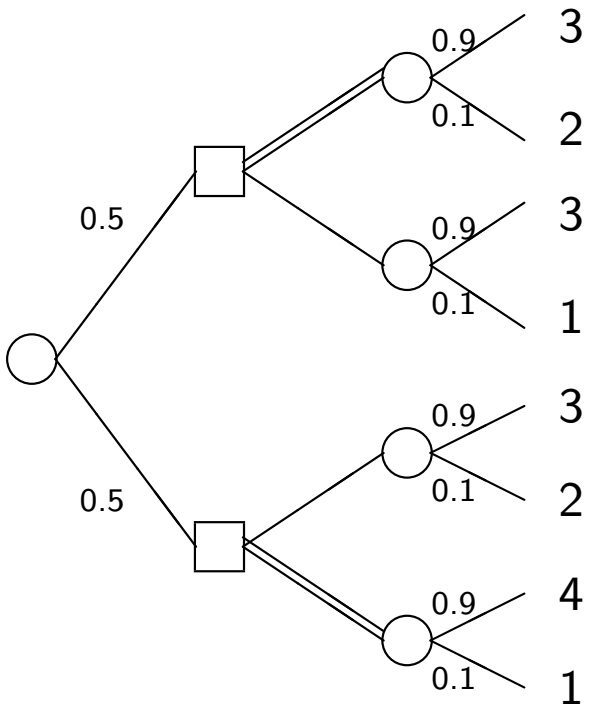


A full problem and the conditional problem "given node 3". The decision problem is time-consistent, if $x_i = \bar{x}_i$, for all nodes, which are in the subtree of the conditioning node.

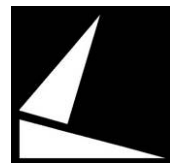


Time inconsistency appears in a natural way in optimality problems: We want to find

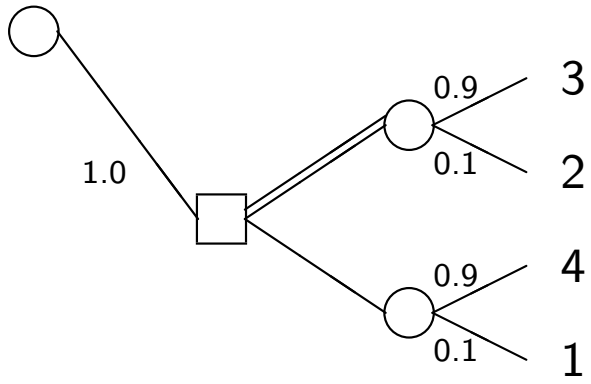
$$\max\{\mathbb{E}(Y) : \mathbb{AV@R}_{0.05}(Y) \geq 2\} \qquad \text{or} \qquad \max \mathbb{E}(Y) + \mathbb{AV@R}_{0.05}(Y).$$



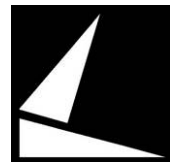
double line = optimal decision



The conditional problem given the first node:



The paradoxon disappears, if the objective is a nested functional, e.g. the nested $\mathbb{A}$V@R or the entropic functional.



Summary about time consistency

- Nested compositions of risk functionals are time consistent (but not interpretable) and information
- Risk functionals applied to the final wealth are typically not time consistent
- Exceptions are only the expectation and the (essential) infimum resp. supremum
- When using time-inconsistent functionals one one has to decide:
  - either to accept time-inconsistent decisions in a rolling horizon setup
  - or to accept decision criteria which depend on the actual path the scenario process takes.

Part II: Pricing of energy contracts

Georg Ch. Pflug

July 3, 2016

General pricing principles: the extremes

- **Insurance pricing.** The contract is seen as exchanging an uncertain position with a certain payment. The pricing operator follows priciples of the insurance business. No optimization is performed.
- **Superreplication pricing.** The risk out of this contract can fully be hedged away by some hedging instruments. The price is the minimal initial capital to do so. It can be found by stochastic optimization.

Insurance premium priciples

If L is the loss variable, there are several ways how insurance companies who cover the full loss L determine the premium price $\pi(L)$ (notice that insurance premia qualify as risk functionals and vice versa):

- **The certainty equivalence principle** (Espen Benth/Cartea/Kiesel 2007): For a disutiliy function V, the premium is
$$\pi(L) = V^{-1}\mathbb{E}_P[V(L)]$$
For $V(x) = \exp(\gamma x)$ this leads to the entropic functional $\pi(L) = \frac{1}{\gamma}\mathbb{E}_P[\exp(\gamma L)]$
- **The change-of measure principle**, e.g. using the Esscher transform (Esscher 1932, Gerber/Shiu 1994, Espen Benth/Sgarra 2009):
$$\pi(L) = \mathbb{E}_Q(L), \text{ where } \frac{dQ}{dP}(x) = c\exp(\gamma x)$$

Positive risk premia are obtained for $\gamma > 0$.

- **The distortion priciple:**

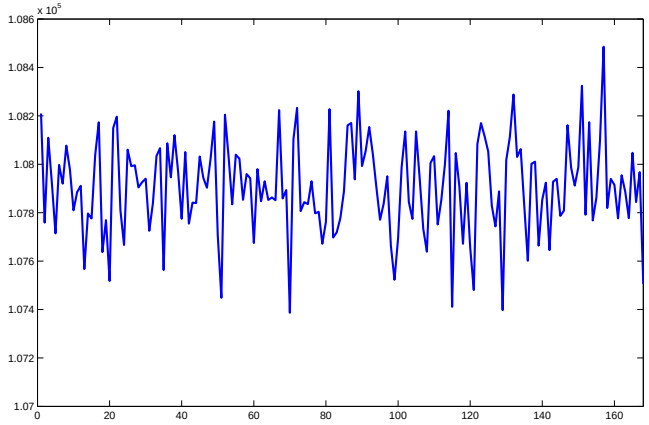
$$\pi(L) = \{\mathbb{E}_Q(L), \text{ with } \frac{dQ}{dP}(x) = h(F_L(x))\} = \int_0^1 F_L^{-1}(p)h(p)\, dp$$

Here h is a distortion function, i.e. a density on [0,1]. It leads to a positive risk premium if h is increasing, and to a negative risk premium if h is decreasing. A typical distortion function is the power distortion

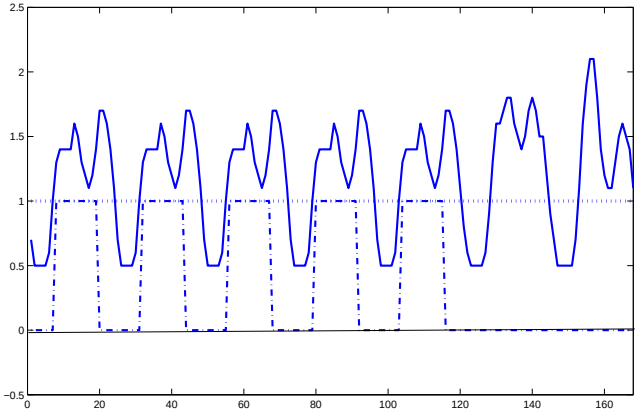
$$h(u) = r(1-u)^{r-1}.$$

$r = 1$ means zero risk premium, $r < 1$ means positive risk premium and $r > 1$ means negative risk premium. This principle coincides with the Esscher transform principle only for exponential distributions.

Demand and hedges: no completeness



Demands



Hedges

Pricing between the extremes

- **No hedging.** Insurance pricing

- **Full hedging.** Superreplication pricing

Pricing between the extremes

- **No hedging.** Insurance pricing
- **Partial hedging.**
  - *Acceptability pricing:* The price is the initial capital needed to hedge some risks under a risk limit for the seller.
  - *Indifference pricing:* The risk limit for the accaptability price is found by considering the risk exposure of the seller before he/she concludes the contract.
  - *Ambiguity pricing:* The model risk is carried by the contract seller.
  - *Bilevel pricing:* The counterparty risk is carried by the contract seller.
- **Full hedging.** Superreplication pricing

Pricing Financial Contracts

- Bid -price: The price of a contract is acceptable for th buyer only if there is no better investment, i.e. no alternative strategy, which for a smaller initial installment would give at least the same outpayments as the given contract.
- Ask-price: The price of the contract is acceptable for the seller only if there is no contract, which pays more at the beginning and has lower or equal liabilities at later stages.

By a duality argument, one sees that the ask-price is always smaller than the bid-price, if there is no arbitrage. In complete markets, they are even equal. Under arbitrage possibilities, the ask-price is larger than the bid-price. The fundamental theorem says that a market is arbitrage free iff there is a probability measure such that the properly discounted price process of the assets is a martingale.

Replication for energy contracts

There is a fundamental difference between pricing in financial markets and pricing in energy markets: While the set of feasible trading strategies is considered to be the same for the seller and for the buyer, it is different in energy markets: The seller has a much larger spectrum of possible actions, he/she may use own energy production, buy energy futures and trade on the wholesale markets, the buyer typically has no access to these possibilities, maybe for the exception of access to the spot markets. For this reason, the seller may determine the upper price by including all his/her assets in a replication model and may offer this price to the buyer.

Notations:

C_t	payments (cash-inflow) to the contract seller
$\mathcal{J} = \{0, \dots, J\}$	energy forms (electricity: $j = 0$, gas, oil, water)
$S^e_{t,j}$	spot prices
$0 \le x^e_{t,j} \le \bar{x}^e_j$	storage and constraints (for electricity $\bar{x}^e_j = 0$)
$x^f_{t,0}$	cash with an interest yield of $r_f > 0$
$y^e_t = (y^e_{t,0}, \dots, y_{t,J})$	amount of energy bought or sold
$d_{t,j} \ge 0$	random inflows (solar, wind, water)
$z^e_{t,ij}$	production of energy i out of energy j
$\underline{z}_{t,ij} \le z^e_{t,ij} \le \bar{z}_{t,ij}$	production limit
η_{ij}	efficiencies of conversion
$\gamma_{t,ij}$	cost factors
$S^f_{t,i}$	financial assets paying cash flows $C^f_{t,i}$
$D_t(t,j)$	delivery of energy j in period $[t, t+1)$ in MWh to th

Equations and constraints

Initialization of energy storages (except electricity)

$$x^e_{0,j} \le x^*_j + y^e_{0,j} + d_{0,j} \tag{1}$$

and

$$x^e_{t,j} \le x^e_{t-1,j} + y^e_{t,j} + \sum_{i=0}^{J} \eta_{ij} z^e_{t-1,ij} - \sum_{i=1}^{J} z^e_{t-1,ji} + d_{t,j} - D_{t,j}. \tag{2}$$

For electricity

$$0 = y^e_{0,0} + d_{0,0} \tag{3}$$

$$0 = y^e_{t,0} + \sum_{i=1}^{J} \eta_{i0} z^e_{t-1,i0} + d_{t,0} - D_{t,0}. \tag{4}$$

Only energy stored at the beginning of a period can be used for conversion during the period:

$$\sum_{j=1}^{J} z^e_{t,ij} \le x^e_{t,i}. \tag{5}$$

Initial cash-account

$$x^f_{0,0} \le w - \sum_{j=0}^{J} S^e_{0,j} y^e_{0,j} - \sum_{i=0}^{I} S^f_{0,i} x^f_{0,i}. \tag{6}$$

and later

$$\begin{aligned} x^f_{t,0} \;\le\;& (1+r_f) x^f_{t-1,0} \\ &- \sum_{j=0}^{J} S^e_{t,j} y^e_{t,j} - \sum_{i=1}^{I} S^f_{t,i}(x^f_{t,i} - x^f_{t-1,i}) + \sum_{i=1}^{I} C^f_{t,i} + C_t \\ &- \sum_{i=0}^{J}\sum_{j=0}^{J} \gamma_{t,ij} z_{t,ij} - \sum_{j=1}^{J} \zeta_j \frac{(x^e_{t,j} + x^e_{t-1,j})}{2} \end{aligned} \tag{7}$$

The terminal inequality ensures that the final asset value is nonnegative

$$x^f_{T,0} + \sum_{j=1}^{J} S^e_{T,j} x^e_{t,j} + \sum_{i=1}^{I} S^f_{T,i} x^f_{T,i} \ge 0. \tag{8}$$

The pricing problem as optimization problem

The (superreplication) price is the minimal value of the following optimization problem:

$$\left\| \begin{array}{l} \text{Minimize (in } x^e, x^f, y, z \text{ and } w) : \ w \\ \text{subject to all constraints} \\ x_t^e, x_t^f, y_t, z_t \text{ are non-anticipative.} \end{array} \right. \tag{9}$$

Acceptability replaces non-replicability

If superreplication leads to unacceptable high or even infinite prices, one may replace the pointwise terminal inequality by

$$\mathcal{U}(x^f_{T,0} + \sum_{j=1}^{J} S^e_{T,j} x^e_{t,j} + \sum_{i=1}^{I} S^f_{T,i} x^f_{T,i}) \geq 0, \qquad (16')$$

where $\mathcal{U}$ is an acceptability functional. In this way, unfavorable scenarios are not avoided completely at the end. Instead, the loss distribution is restricted by the acceptability functional, such that only unfavorable outcomes with small probability are acceptable.

The resulting optimization problem for acceptability pricing is a modification and can be written as

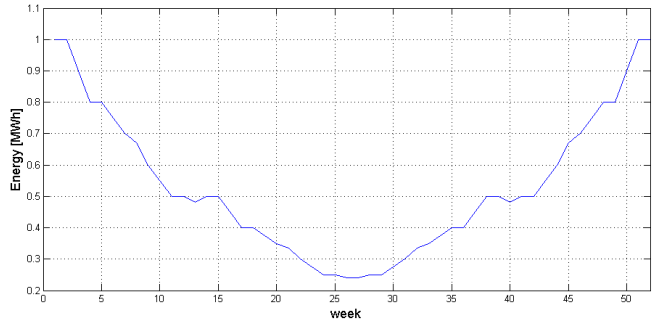
$$\left\| \begin{array}{l} \text{Minimize (in } x^e, x^f, y, z \text{ and } w) : w \\ \text{subject to the given constraints} \\ \mathcal{U}(x^f_{T,0} + \sum_{j=1}^{J} S^e_{T,j} x^e_{t,j} + \sum_{i=1}^{I} S^f_{T,i} x^f_{T,i}) \geq 0 \\ x^e_t, x^f_t, y_t, z_t \text{ are non-anticipative.} \end{array} \right. \tag{10}$$

An example

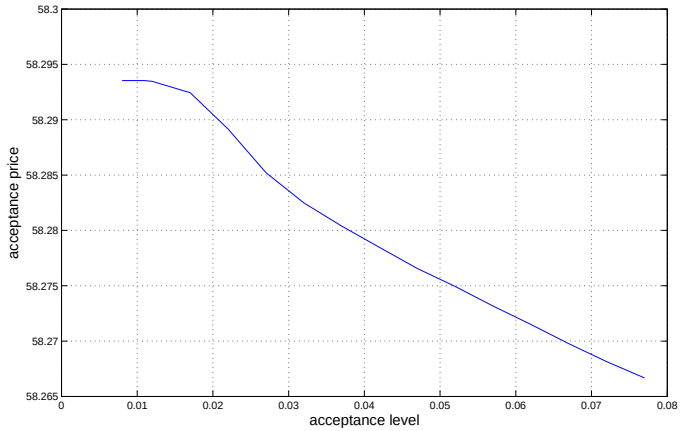
We consider a planning horizon of one year (52 weeks). Electricity spot prices are modeled by geometric Brownian motion with jumps (GBMJ), estimated from EEX Phelix hourly electricity prices (hourly, 09/2008-12/2011, Bloomberg). The pricing model was reformulated and solved on a stochastic tree, generated from the GBMJ model.

The hedging opportunities are represented by four futures contracts, related to the quarters of the year, i.e. each of the futures delivers a constant amount of electric energy during one of the quarters.

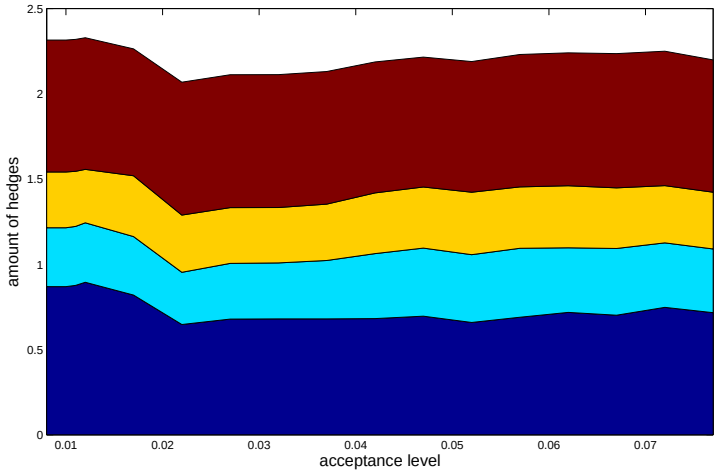
By solving the stochastic optimization problem with $\mathbb{A}$V@R-objective, the acceptability price is calculated for a pure trader meaning that only wholesale base quarter future contracts can be used for hedging for different values of the $\mathbb{A}$V@R-parameter α.



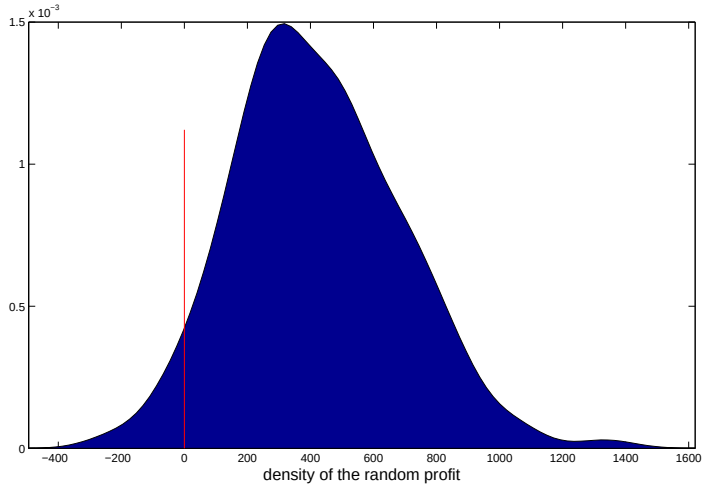
Acceptability pricing: delivery pattern D_t over 52 weeks.



Acceptability pricing: The price of 1 MWh as a function of the acceptance level α.



Acceptability pricing: optimal hedges as a function of the acceptance level α.



Acceptability pricing: density of the profit variable

Acceptability pricing for financial contracts

For purely financial contracts, the acceptability upper pricing problem can be investigated in more detail: The upper price π_u is the minimal value of

$$\left\| \begin{array}{l} \text{Minimize (in } x \text{ and } w) : w \\ \text{subject to} \\ Y_0^{w,x} = w \\ Y_{t-}^{w,x} - Y_t^{w,x} \geq c_t \qquad t = 1, \ldots, T \\ \mathcal{U}(Y_T^{w,x}) \geq 0 \\ x_t \text{ is nonanticipative} \end{array} \right. \tag{11}$$

which takes the following form in the linear setup:

$$\left\| \begin{array}{l} \text{Minimize (in } x \text{ and } w) \; w \\ \text{subject to} \\ x_0 S_0 \leq 0 \\ x_{t-1} S_t - x_t S_t - c_t \geq 0; t = 1, \ldots, T \\ \mathcal{U}_T(x_T S_T) \geq 0 \end{array} \right. \tag{12}$$

If the functional $\mathcal{U}$ is given by its dual representation $\mathcal{U}(Y) = \inf\{\mathbb{E}(YZ) : Z \in \mathcal{Z}\}$, then the acceptability pricing problem has a dual:

$$\left\| \begin{array}{l} \text{Maximize (in } Z_t) \sum_{t=1}^{T} \mathbb{E}(c_t\, Z_t) \\ \text{subject to} \\ \mathbb{E}(S_{t+1} Z_{t+1} | \mathcal{F}_t) = Z_t S_t \\ Z_t \geq 0; t = 1, \ldots, T-1 \\ Z_T \in \mathcal{Z} \end{array} \right. \tag{13}$$

The latter problem can be reformulated as before: Let $\tilde{c}_t = c_t / S_{t,0}$ and $\tilde{S}_t = S_t / S_{t,0}$, where $S_{t,0}$ is the riskless investment. Then the acceptability upper price π_u is given by

$$\max\{\sum_{t=1}^{T} \mathbb{E}_Q(\tilde{c}_t) : \ (\tilde{S}_t) \text{ is an (equ.) martingale under } Q \text{ s. t. } \frac{dQ}{dP} \in \mathcal{Z}\} \tag{14}$$

Similarly, the lower price π_ℓ is

$$\min\{\sum_{t=1}^{T} \mathbb{E}_Q(\tilde{c}_t) : \ (\tilde{S}_t) \text{ is an (equ.) martingale under } Q \text{ s. t. } \frac{dQ}{dP} \in \mathcal{Z}\}$$

Denote by $\pi_u(\mathcal{Z})$ the upper price in dependency of the considered acceptability functional $\mathcal{U}$ with dual set $\mathcal{Z}$. Notice that

$$\mathcal{Z}_1 \subseteq \mathcal{Z}_2 \quad \text{implies that} \quad \pi_u(\mathcal{Z}_1) \leq \pi_u(\mathcal{Z}_2).$$

The largest price is gotten when full (super)replication is required, meaning that $\mathcal{Z}$ must be equal to all nonnegative random variables. A smaller and more realistic price is obtained, if the acceptability functional is e.g. the average value-at-risk $\mathbb{A}\text{V@R}_\alpha$, with $\mathcal{Z} = \{Z : 0 \leq Z \leq \frac{1}{\alpha}\}$. The smallest price is given by the choice $\mathcal{Z} = \{1\}$, which corresponds to the acceptability requirement $\mathbb{E}(Y_T^{w,x}) \geq 0$. This simple pricing rule is related to the concept of expected net present value (ENPV) of a contract and can be seen as the absolute minimum price for avoiding bankruptcy. However no seller will be willing to contract on this basis.

Indifference pricing

The indifference principle states that the seller of a product compares his optimal decisions with and without the contract and then requests a price such that he is at least not worse off when closing the contract. This idea goes back to insurance mathematics (Buehlmann, 1972) but has been used for pricing a wide diversity of financial contracts in recent years, see Carmona (2009) for an overview.

In order to model the indifference price approach, assume that the total energy deliveries of the actual portfolio are D_t^{old} and the total cash-flows out of this portfolio are C_t^{old}. These cash-flows must include also the upfront payments at time 0. The additional contract, for which a price is not yet determined, is given by D_t resp. C_t.

Indifference pricing happens in two steps:

- Determination of the acceptability of the actual portfolio. To this end, the following problem is solved:

$$\left\| \begin{array}{l} \text{Maximize (in } x,y,z \text{ and } w) : \mathcal{U}(C_T^{old} + \sum_{j=1}^{J} S_{T,j}^e x_{t,j}^e + \sum_{i=1}^{I} S_{T,i}^f x_{T,} \\ \text{subject to the constraints} \\ x_t, y_t, z_t \text{ are non-anticipative.} \end{array} \right. \tag{16}$$

  Here the equations are based on D_t^{old} resp. C_t^{old}. The optimal value of this optimization problem, that is the acceptability level of the actual (old) portfolio, is denoted by a_0.
- Determination of the indifference price of the additional contract. Let the new total deliveries be $D_t^{new} = D_t^{old} + D_t$ and the new cash-flows (without the upfront payment for the additional contract) be $C_t^{new} = C_t^{old} + C_t$. The price of the additional contract is denoted by $x_{0,0}$. It is determined by the following problem:

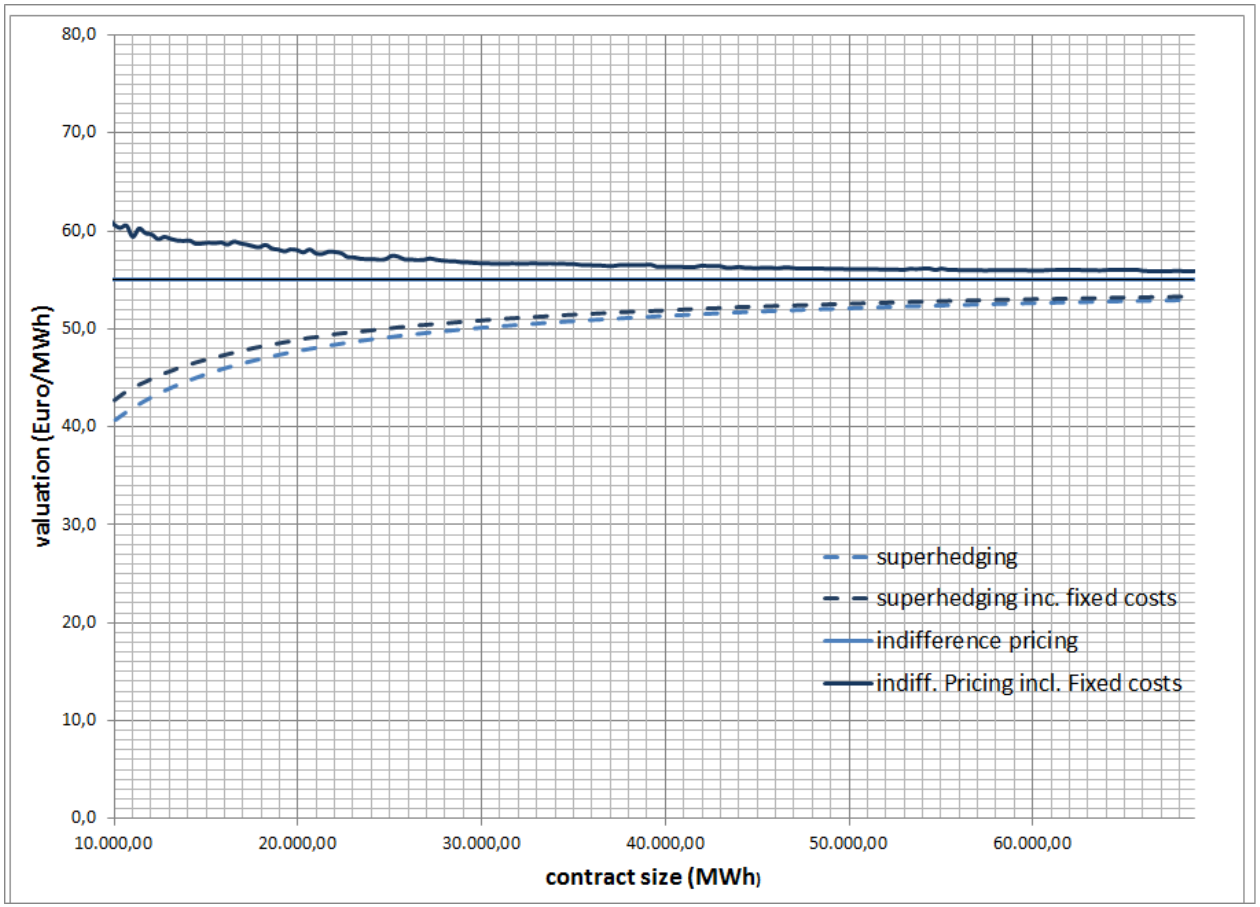
$$\left\| \begin{array}{l} \text{Minimize (in } x,y,z \text{ and } w) : w \\ \text{subject to} \end{array} \right.$$

Example

Consider an electricity producer, who has available a single combined cycle plant that is able to use both, oil and gas. The machine has maximum power production of 410 MW and efficiencies of 0.575 (gas) and 0.57 (oil). Both fuels can be stored up to some amount ($1.5 \cdot 10^6$ MWh) at storage costs 0.2 Euro/MWh/h. We do not consider futures contracts in this setup, hence hedging is possible only by buying fuel at appropriate points in time. Again we use electricity prices and weekly decision periods. We valuate a simple delivery contract, which binds the producer to supply a fixed amount of energy, the contract size in MWh, during each stage of the planning problem. The producer is free to buy and store fuel, and to produce electric energy for the contract and also for selling it at the spot market. The value of the contract per MWh contains variable operating costs. From this we calculate a contract value per MWh that also includes an amount of coverage for fixed cost, which is proportional to the mean workload of the production unit during the planning horizon.

Within this setup we compare the superhedging approach to indifference pricing. Superhedging is possible for a producer, if the contract size does not exceed the capacity of the combined cycle turbine. For indifference pricing, we use the average value-at-risk at level $\alpha = 0.05$ as the related acceptability functional.
As described above, superhedging leads to contractual deliveries, but does not account for alternative usages of the machine, whereas indifference pricing does. This is the reason, why the superhedging price might be considered as too low in this case. Superhedging and indifference pricing also show different amounts of fixed cost, because the related strategies use non-contractual electricity production at different levels.

Superhedging and indifference pricing



Electricity swing options

- Give the buyer the right to get electricity at a predetermined price K per MWh during the contract period (a month, a quarter, a year, ...). The price is set way before the delivery period starts.
- The actual schedule of hourly demand y_t may deviate from some baseline schedule: Usually the cumulative demand must lie in some interval $[\underline{E}, \overline{E}]$ and the demand in hour t must lie within $[\underline{e}_t, \overline{e}_t]$.
- The buyer has to announce its actual demand for each hour with 1 day ahead notice.
- The seller can use forward contracts, own production and the spot market to meet the demand of the buyer.
- The buyer can use the spot market as an alternative for his demand.
- The problem is to determine a fair offered price K for the contract.

Bilevel problems

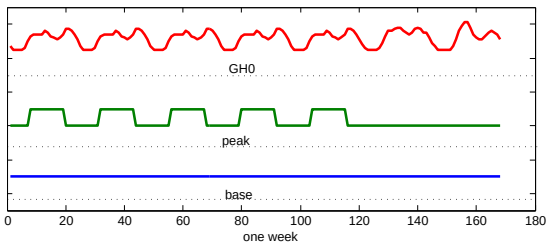
In bilevel problems, there are two independent decision makers (DMs): The typical situation is that the upper level DM sets the price of a contract while the lower level DM decides about to accept the contract and to exercise the rights out of the contract.

In stochastic bilevel problems, there are random parameters, which are only known by their distribution.

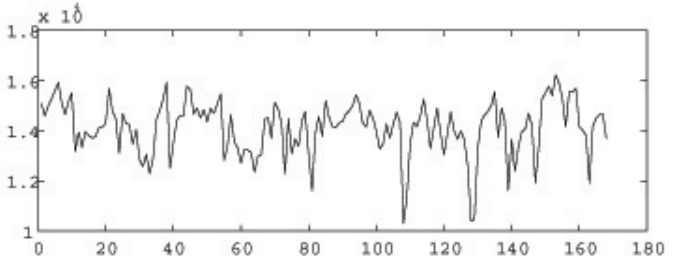
Flexible electricity contracts: Upper level decisions

The upper level DM (contract seller) has to decide about the offered price for the contract as well as about production and the hedges to buy from future markets. Examples of available hedging profiles:

The future market offers hedges with given profiles:



However, the contract buyer has an irregular demand:



Full hedge is not possible: A basic risk remains with the contract seller.

There are M hedging instruments available. A hedging instrument is characterized by its price F_m and its delivery pattern τ, where $\tau(m,t)$ is the amount delivered in time period (hour) t.
The decision to be made by the option seller consists of

- the number $\tilde{x}_m$ of units of future contract m to be bought,
- the ask price K.

We collect the hedge amounts in a hedge vector $\tilde{x} = (\tilde{x}_1, \ldots, \tilde{x}_M)$ and all upper level decision variables in

$$x = (\tilde{x}, K).$$

The unmatched surplus/shortage in time period t is

$$\sum_{m=1}^{M} \tilde{x}_m \tau(m,t) - y_t.$$

This amount will be sold/bought on the spot market. The spot-prices are random processes S_t with a given distribution. The input data of the optimal hedging problem are

$S = (S_t^\omega)$	the spot price scenario model
$p = (p^\omega)$	the scenario probabilities
$F = (F_m)$	the prices of the hedging instruments (future contracts)
$\tau(m,t)$	the delivery pattern of hedge m
$y = (y_t^\omega)$	the demands (which are decided by the LL DM)

The revenue as a function of $x = (\tilde{x}, K)$ and y

For consistency reasons, we assume that the spot price model is calibrated to the known future prices F_m is such a way, that *expectation-neutrality* holds: $F_m = \sum_{t=1}^{T} \tau(m,t)\mathbb{E}[\xi_t]$
Denote by W the profit/loss variable of this contract seen from the option seller. Given the hedge $\tilde{x}$ and the price per unit K, W takes the values

$$W_x^\omega = \bar{y}K - \delta(y) + \sum_{m=1}^{M} \tilde{x}_m[\phi_m - F_m]$$

with probability p^ω. Here $\bar{y} = \sum_{t=0}^{T} y_t$ is the total demand, $\delta(y) = \sum_{t=0}^{T-1} y_t S_t$ is the spot-price value of the demand and $\phi_m = \sum_t S_t \tau(m,t)$ is the spot-price value of one unit of hedge m.

Monopolistic case: Maximizing expected profit

$$[UL_M] \left| \begin{array}{l} \max_{K,\tilde{x},y} \mathbb{E}[W_x] = \mathbb{E}[\bar{y}K - \delta(y) + \sum_{m=1}^{M} \tilde{x}_m[\phi_m - F_m]]. \\ \text{subject to} \\ \sum_{t=0}^{T-1} \mathbb{E}\left[y_t\left(S_{t+1} - K\right)\right] \geq 0 \\ y \in \Psi(x) \end{array} \right. \tag{18}$$

where

$$[LL] \left| \begin{array}{l} \Psi(x) = \text{argmax}_{\ y} \ \sum_{t=0}^{T-1} \mathbb{E}\left[y_t\left(S_{t+1} - K\right)\right] \\ \text{subject to} \\ \underline{e}_t \leq y_t \leq \overline{e}_t, \ \forall t \in \{0, \ldots, T-1\} \\ \underline{E} \leq \sum_{t=0}^{T-1} y_t \leq \overline{E} \\ y_t \geq 0, \quad \forall t \in \mathcal{T} \\ y_t \triangleleft \mathcal{G}_t, \end{array} \right. \tag{19}$$

Competitive case: Minimizing the price under an acceptability condition

$$[UL_C] \left| \begin{array}{l} \min_{K,\tilde{x},y} K \\ \text{ subject to} \\ \mathcal{U}[W_x] = \mathcal{U}[\bar{y}K + \sum_{m=1}^{M} \tilde{x}_m[\phi_m - F_m] - \delta(y)] \geq q \\ \sum_{t=0}^{T-1} \mathbb{E}\left[y_t\left(S_{t+1} - K\right)\right] \geq 0 \\ y \in \Psi(x), \end{array} \right. \quad (20)$$

where

$$[LL] \left| \begin{array}{l} \Psi(x) = \text{argmax}_{\ y} \ \sum_{t=0}^{T-1} \mathbb{E}\left[y_t\left(S_{t+1} - K\right)\right] \\ \text{ subject to} \\ \underline{e}_t \leq y_t \leq \overline{e}_t, \ \forall t \in \{0, \ldots, T-1\} \\ \underline{E} \leq \sum_{t=0}^{T-1} y_t \leq \overline{E} \\ y_t \geq 0, \quad \forall t \in \mathcal{T} \\ y_t \triangleleft \mathcal{G}_t, \end{array} \right. \quad (21)$$

Reformulation

Both decision problems (with $\mathcal{A} = \mathbb{A}\mathrm{V@R}_\alpha$) are reformulated with finite state space. This is done by representing the filtration by a tree structure and and relating all relevant values (prices, probabilities and decisions) to the nodes. After choosing appropriate matrices and vectors we can use the reformulation.

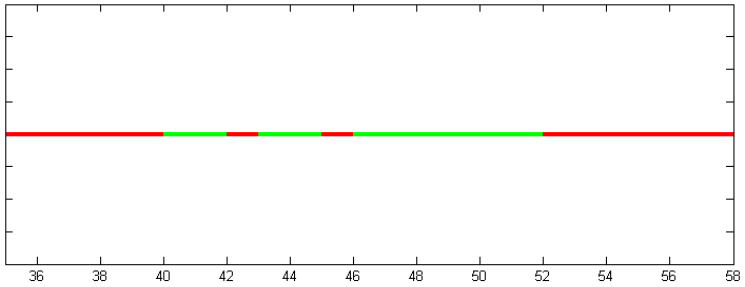
$$\begin{aligned}
&\text{maximize (in } (x,y))\, x^\top (d + Dy) \\
&Ex \geq \ell \\
&y \in \Psi(x)
\end{aligned}$$

$$\begin{aligned}
&\text{maximize (in } (x,y))\, x^\top (d + Dy) \\
&q_x^{i\top} x + q_y^{i\top} y + x^\top Q^i y \geq r^i \quad i = 1, \ldots, n_r \\
&Ex \geq \ell \\
&y \in \Psi(x)
\end{aligned}$$

where $\Psi(x) = \operatorname{argmax} \; (c^\top + x^\top C)y$ subject to $Ay \leq b$

Acceptability conditions in the LL

If the LL contains some acceptability conditions (e.g. that the expected profit is nonnegative), the LL problem may become unfeasible if the strike price K is too high. On the other hand, the UL problem may become infeasible if the strike price K is too low. Thus the total feasibility region lies between an lower and an upper bound (but may be a union of non-intersecting intervals). In the competitive case, we search for the lowest feasible point.



MPEC reformulation

- Most popular approach: Instead of the bilevel problem solve a problem where $y \in \Psi(x)$ is replaced by the KKT conditions of the lower level problem.
- Consider a bilevel problem as defined before and assume that for each $x \in X$ the lower level problem is convex and fulfills the MFCQ at each feasible point. Then any local optimal solution of the bilevel problem is a local optimal solution of the related MPEC reformulation.
- For the competitive case this leads to

$$
\begin{aligned}
&\text{maximize (in } x, y, \lambda) x^\top (d + Dy) \\
&q_x^{i\top} x + q_y^{i\top} y + x^\top Q^i y \geq r^i \qquad i = 1, \ldots, n_r \\
&Ex \geq \ell \\
&c^\top + x^\top C = \lambda^\top A \\
&Ay \leq b \\
&\lambda^\top (Ay - b) = 0 \\
&\lambda \geq 0
\end{aligned}
$$

- This MPEC-formulation is also hard to solve because of the included complementarity constraints.

Necessary optimality conditions

- Mathematical programs with equilibrium constraints can be written in the form

$$\begin{aligned} &\min_x f(x) \qquad (22)\\ &s.t. \quad C^E(x) = 0 \quad C^I(x) \geq 0\\ &\qquad G(x) \geq 0 \quad H(x) \geq 0\\ &\qquad G(x)^T H(x) = 0 \end{aligned}$$

- One difficulty of this nonconvex class of optimization problems arises from the fact that usual constraint qualifications fail, and KKT-type conditions therefore are more involved.
  - MFCQ are violated in every feasible point.
  - This is true also for the stronger LICQ and Slater conditions

Necessary optimality conditions

- It is possible to use the KKT conditions for (22), if the Guignard constraint qualification (GCQ) is fulfilled.
- Let the feasible set of an optimization problem be described by $S = \{x : \ C^E(x) = 0, \ C^I(x) \geq 0\}$ and $\mathcal{I}$ denote the sets of binding inequality constraints. The cone of tangents of the feasible set is given by

$$T_S(x_0) = \{d : \ d = \lim_{k \to \infty} \lambda_k (x_k - x_0), \ \lambda_k > 0, \ x_k \in S \to x_0\}$$

and the linearized tangent cone is defined as

$$L'_S(x_0) = \left\{ \ d : \quad \nabla C_i(x)^T d = 0, \ i \in E, \ \nabla C_i(x)^T d \geq 0, \ i \in \mathcal{I} \ \right\}.$$

- If x_0 is a local optimum of the optimization problem, then x_0 is B-stationary, i.e.

$$\nabla f(x_0) \in T_S(x_0)^*$$

- If the GCQ

$$T_S(x_0)^* = L'_S(x_0)^*$$

holds, then x_0 is B-stationary if and only if it is linearized B-stationary, i.e. $\nabla f(x_0) \in L'(x_0)^*$.

- In this case any optimal point x_0 fulfills the (usual) KKT conditions. The Lagrange multipliers will not be unique.
- It may be hard to calculate the tangent cone and its dual.

Strong stationarity

- A feasible point x_0 of the MPEC (22) is called *strongly stationary* if it is a KKT point of the modified (localized, relaxed) problem

$$\begin{aligned} &\min_x f(x) && (23)\\ &s.t. \quad c^E(x) = 0 \quad c^I(x) \geq 0\\ &\qquad G_i(x) = 0 \quad \text{if } H_i(x_0) > 0\\ &\qquad G_i(x) \geq 0 \quad \text{if } H_i(x_0) = 0\\ &\qquad H_i(x) = 0 \quad \text{if } G_i(x_0) > 0\\ &\qquad H_i(x) \geq 0 \quad \text{if } G_i(x_0) = 0 \end{aligned}$$

- If x_0 is a solution of the MPEC and the GCQ holds at x_0, then x_0 is strongly stationary.
- MPEC-LICQ: If the LICQ holds for (23) then strong stationarity again is a necessary optimality condition. The related Lagrange multipliers are unique.

Strong stationarity for the competitive swing option problem

- A feasible solution (x, y, λ) of the MPEC reformulation is a strongly stationary solution if there exist multipliers μ, ν_1, ν_2 such that the following conditions are fulfilled:

1. $d^\top + y^\top D = \sum_{i=1}^{n} \mu_{1i}(q^{i\top} + y^\top Q^{i\top}) + \mu_2^\top E - \mu_3^\top C^\top$
2. $x^\top D = \sum_{i=1}^{n_r} \mu_{1i}(q^{i\top} + x^\top Q^i) - \nu_2^\top A$
3. $\mu_3^\top A^\top + \nu_1^\top = 0$
4. $\mu_1 \geq 0,\ \mu_2 \geq 0$
5. $\sum_{i=1}^{n} \mu_{1i} \left(q_x^{i\top} x + q_y^{i\top} y + x^\top Q^i y - r^i\right) = 0$
6. $\mu_2^\top \cdot (Ex - \ell) = 0$
7. $\nu_{1j} \geq 0$, if $\lambda_j = 0$ and $A_{j\cdot} y = b_j$
8. $\nu_{1j} = 0$ if $\lambda_j > 0$ and $A_{j\cdot} y = b_j$
9. $\nu_{2j} \geq 0$, if $\lambda_j = 0$ and $A_{j\cdot} y = b_j$
10. $\nu_{2j} = 0$ if $\lambda_j = 0$ and $A_{j\cdot} y < b_j$

MPEC-LICQ for the swing option problem

- *Let*
  - $\hat{E}$ *denote the matrix of those rows* e_i *of* E *with* $e_i x = \ell_i$.
  - $\hat{Q}_x(y)$ *contain all rows* $\left[q_x^{i\top} + y^\top Q^{i\top}\right]$ *and* $\hat{Q}_y(x)$ *contain all rows* $\left[q_y^{i\top} + x^\top Q^i\right]$, *where* $q_x^{i\top} x + q_y^{i\top} y + x^\top Q^i y = r^i$.
  - $\hat{A}$ *denote the matrix of rows* a_i *fulfilling* $a_i \cdot y = b_i$.
  - $\hat{I}$ *denote the matrix of rows* $e_i^\top$ *(the i-th unit vectors), where* $\lambda_i = 0$
- *Then for any feasible* (x, y) *and* $\lambda \in M(x, y)$ *the MPEC-LICQ holds at* (x, y) *if the matrix*

$$\begin{bmatrix} C^\top & 0 & -A^\top \\ \hat{Q}_x(y) & \hat{Q}_y(x) & 0 \\ \hat{E} & 0 & 0 \\ 0 & \hat{A} & 0 \\ 0 & 0 & \hat{I} \end{bmatrix}$$

has full row rank.

M-stationarity for the swing option problem

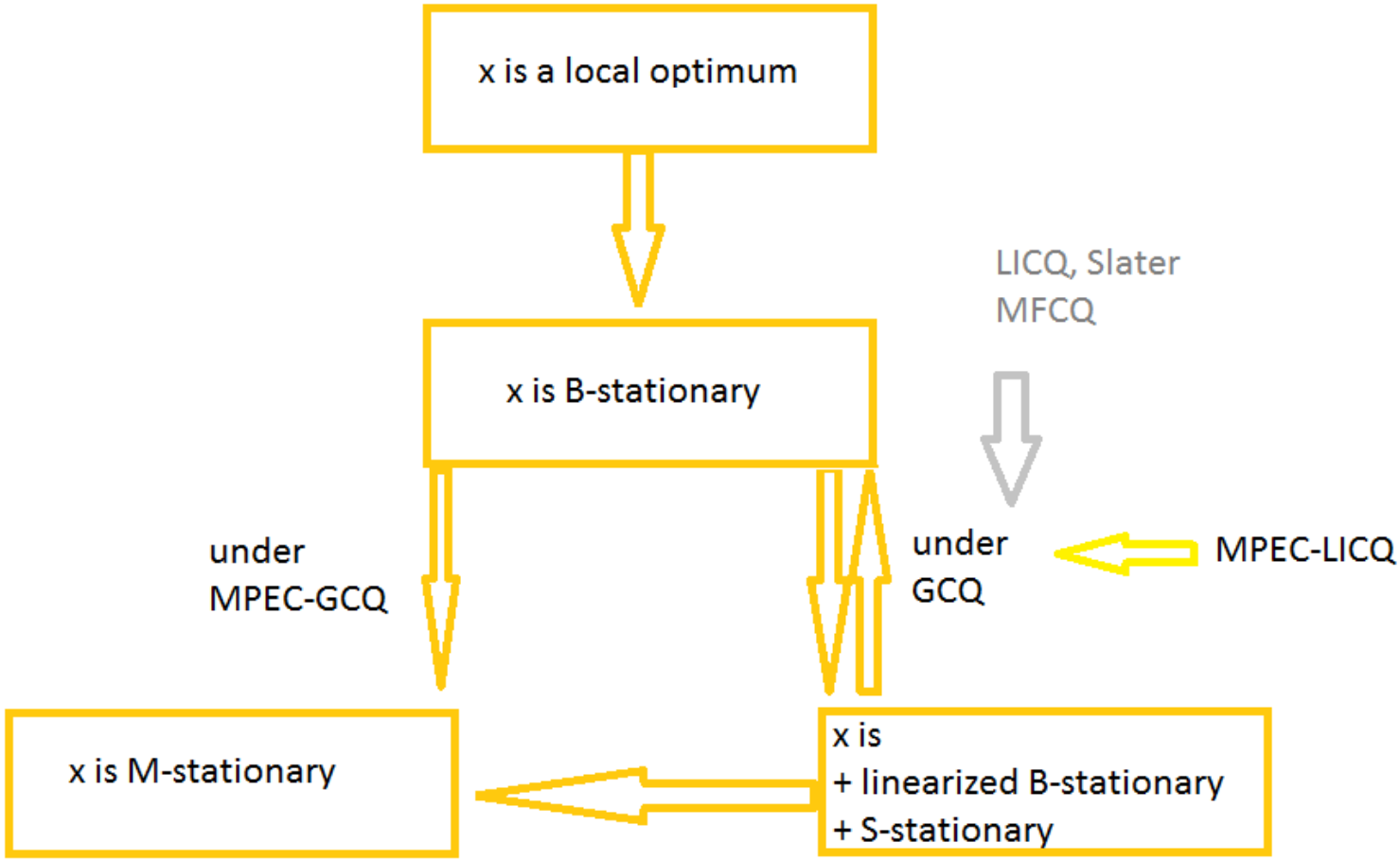
- It is possible that for an MPEC the MPEC-LICP and even the weaker GCQ are violated.
- A good alternative necessary condition can be M-stationarity under a modified Guignard-type constraint qualification. (Flegel, Kanzow, Outrata 2007). This idea is based on Mordukhovich cones.
- M-stationarity: Conditions 7-10 for strong stationarity are replaced by
  1. $\nu_{1j} = 0$, if $\lambda_j > 0$ and $A_{j\cdot} y = b_j$
  2. $\nu_{2j} = 0$, if $\lambda_j = 0$ and $A_{j\cdot} y < b_j$
  3. $\nu_{1j} > 0 \wedge \nu_{2j} > 0$ or $\nu_{1j} = 0 \wedge \nu_{2j} = 0$, if $\lambda_j = 0$ and $A_{j\cdot} y = b_j$
- The MPEC-linearized cone $\mathcal{T}^{lin}_{MPEC}$ is similar to the linearized cone L'_S, but contains the additional constraints

$$\left(\nabla G_i(x_0)^\top d\right)\left(\nabla H_i(x_0)^\top d\right) = 0, \text{ if } G_i(x_0) = H_i(x_0) = 0$$

- The MPEC-GCQ then can be written as

$$T(x_0)^* = \mathcal{T}^{lin*}_{MPEC}$$

Necessary conditions



Algorithms: Using the MPEC-reformulation

- Based on the MPEC reformulation several approaches could be tried to solve the swing option problem
  - The MPEC formulation can be again reformulated to a (binary) mixed integer problem, discriminating between the cases $\lambda_i > 0$ and $\lambda_i = 0$
    - For the swing option problem there are many integer variables for reasonable instances
    - Some bilinear constraints still remain
  - Inner point algorithms are not applicable directly
    - However, several methods for relaxing the complementarity constraints or for penalizing them were proposed in literature.
  - SQP works fine (under some additional regularity conditions) for strongly stationary optimal points, which fulfill the MPEC-LICQ.
- We developed also some algorithms that build on special features of the swing option problem.

Monopolistic case: dual gap reformulation

- The upper level can be formulated as an LP, while given the upper level decisions x the lower level can be formulated also as an LP. Consider e.g. the lower level

$$[LL-P] \quad \left| \begin{array}{l} \max_y (c^\top + x^\top C)y \\ Ay \leq b \\ y \geq 0 \end{array} \right.$$

for some suitable vectors b, c_1 and matrices A, C_2.with dual

$$[LL-D] \quad \left| \begin{array}{l} \min_v b^\top v \\ A^\top v - C^\top x \geq c \\ v \geq 0 \end{array} \right.$$

- Minimizing the duality gap is equivalent to solving the lower level problem, which can be used for the following penalized version of the bilevel problem.

$$[ULM+LL] \quad \left| \begin{array}{l} \max_{x,v;y} x^\top (d + Dy) + \gamma[(c + x^\top C)y - b^\top v] \\ Ex \geq \ell, Ay \leq b, A^\top v - Cx \geq c \\ y \geq 0, v \geq 0. \end{array} \right.$$

Monopolistic case: Algorithm

1. Set $K := K_0$, as starting value
2. Solve [ULM+LL] in $x = (K, \tilde{x})$ with K fixed, resulting in the solutions $\tilde{x}$, v, y
3. Repeat, until $|K - K_{old}| \leq \epsilon$
   - 3.1 Set $K_{old} := K$
   - 3.2 Solve [ULM+LL] with y fixed resulting in the solutions K, $\tilde{x}$ and v
   - 3.3 Solve [ULM+LL] with x, v fixed resulting in y
4. Set $K^* := K$, $x^* := x$, $y^* := y$

Competitive case: Algorithm

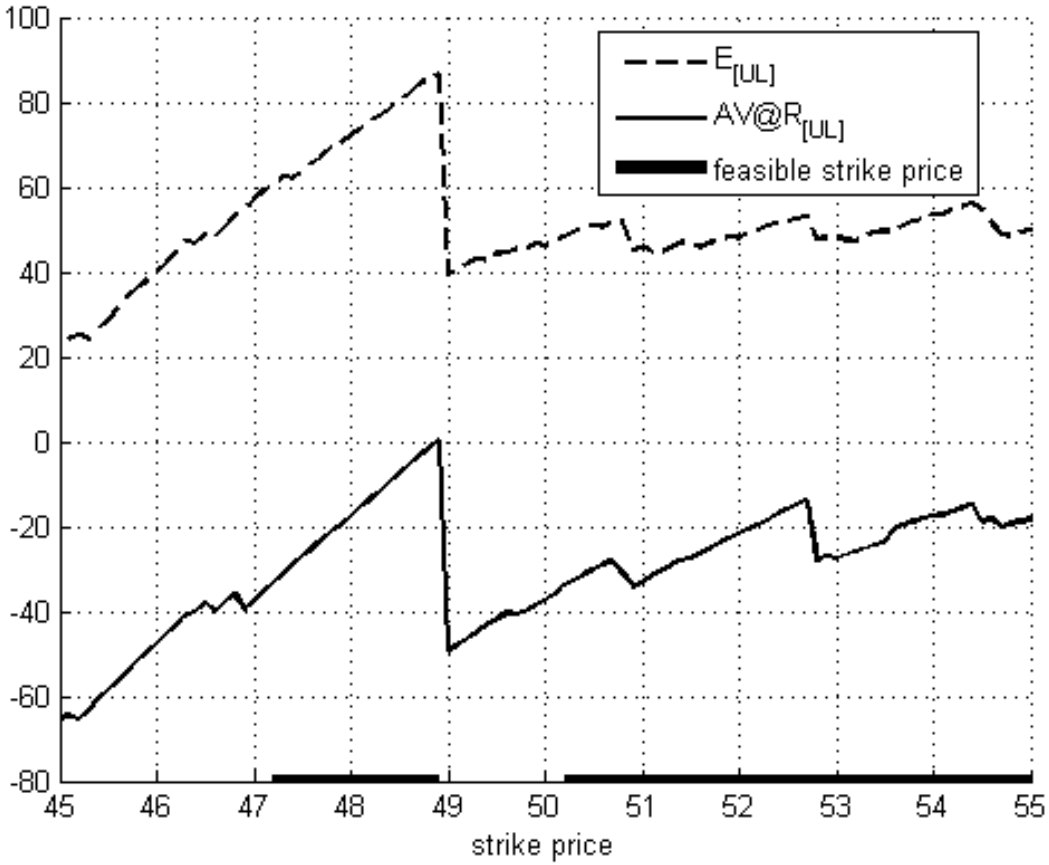
- A similar dual-gap reformulation can be stated for the competitive problem.
- However, the above algorithm can not be applied in this case (bilinear constraints).A second simple algorithm uses the fact that the strike price K is the main decision variable. A bracketing type procedure is used to find the lowest level of K such that all constraints are satisfied, and the dual gap of the lower level problem is closed.

Competitive case: Algorithm

Set $l_{bound} := 0$, $u_{bound} := 2K$, $\delta := u_{bound} - l_{bound}$, $K^* := \infty$
While $\delta > \epsilon_1$

1. Set $K := \frac{(u_{bound} - l_{bound})}{2}$
2. Solve $[UL_C + LL_E]$ with K fixed, set $feas(UL)$=true if it is feasible(and otherwise false)
3. If $feas(UL)$
   - 3.1 Get the solutions $\tilde{x}$, y, v
   - 3.2 Set the duality gap $\eta := (c_1 + x^\top C_2)y - b^\top v$ for $x = (K, \tilde{x})$
4. EndIf
5. If $|\eta| > \epsilon_2$ or $feas(UL)$=false
   - 5.1 Solve [LL_E] and set $feas(LL) = true$ if it is feasible (and otherwise false)
   - 5.2 If $feas(LL) =$ false}
     - 5.2.1 $u_{bound} := K$
   - 5.3 Else
     - 5.3.1 $l_{bound} := K$
   - 5.4 EndIf
6. Else
   - 6.1 Set $K^* := K$, $\tilde{x}^* := \tilde{x}$, $y^* := y$
7. EndIf
8. Set $\delta := u_{bound} - l_{bound}$
9. EndWhile

The solution of the competitive case

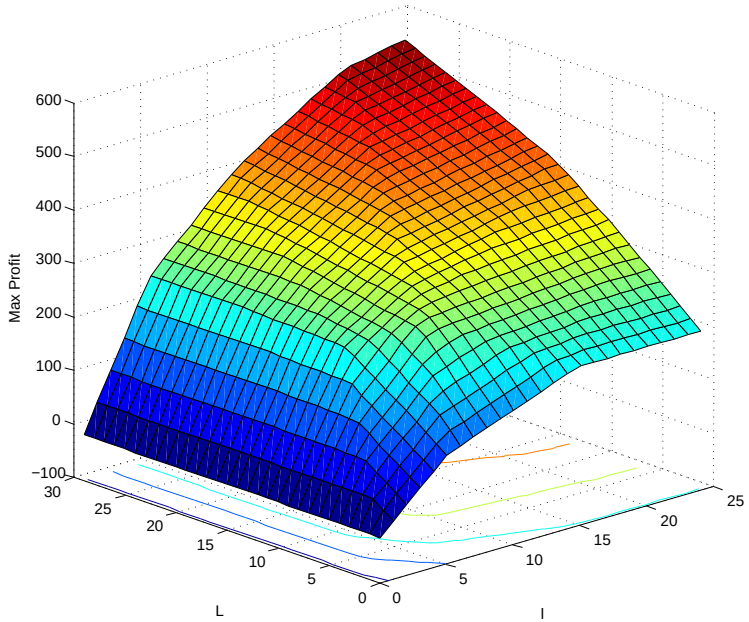


The optimal UL decision is the lowest point of the feasible region.

Hedging

The optimal hedges depend on the strike price K

The costs for flexibility



Flexibility for the contract buyer = costs for the contract seller

Conclusions

Stochastic bilevel programs add an additional complexity to deterministic bilevel programs (which are already hard problems). The swing option problem has the following peculiarities:

- The correct pricing of a swing option requires to find a behavioral model for the contract holder, in particular his price sensitivity (full reseller, partial reseller, no reseller).
- A worst case can be found by considering the optimizing strategy for the buyer assuming he is a reseller (as we did it).
- The optimal hedging strategies for fixed contracts and swing options may be quite different.
- Higher flexibility of requires a higher price (the costs of flexibility).

Part III: Optimal hydro management and the ambiguity problem

Georg Ch. Pflug

July 3, 2016

The ambiguity problem

Let the basic problem be

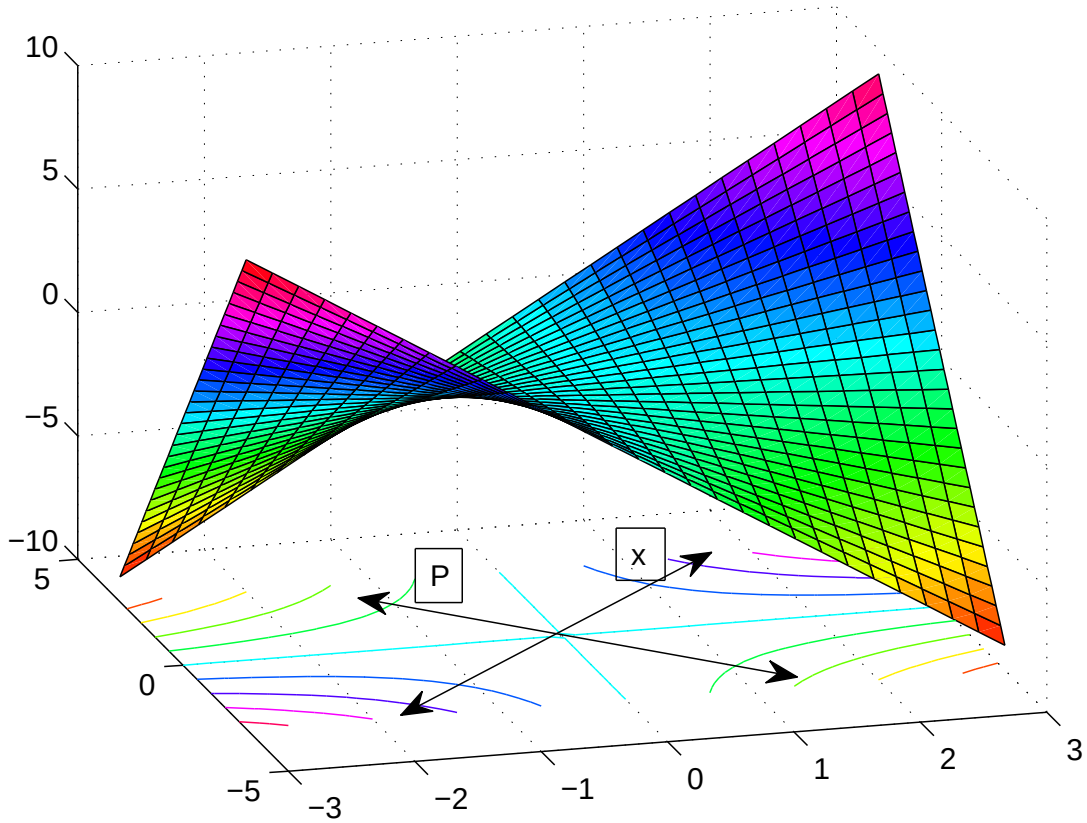
$$\min\left\{\mathbb{E}_{\hat{P}}[Q(x,\xi)] \,:\, x\in\mathbb{X}\right\}$$

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 $P^*\in\mathcal{P}$.

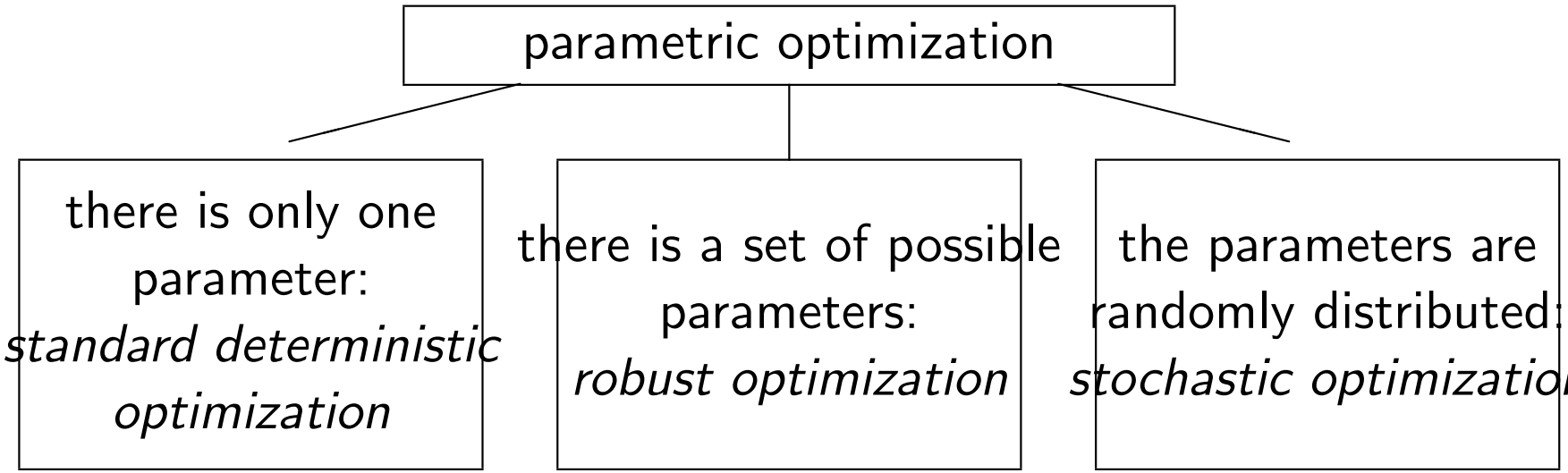
The pair (x^*, P^*) forms a saddle point

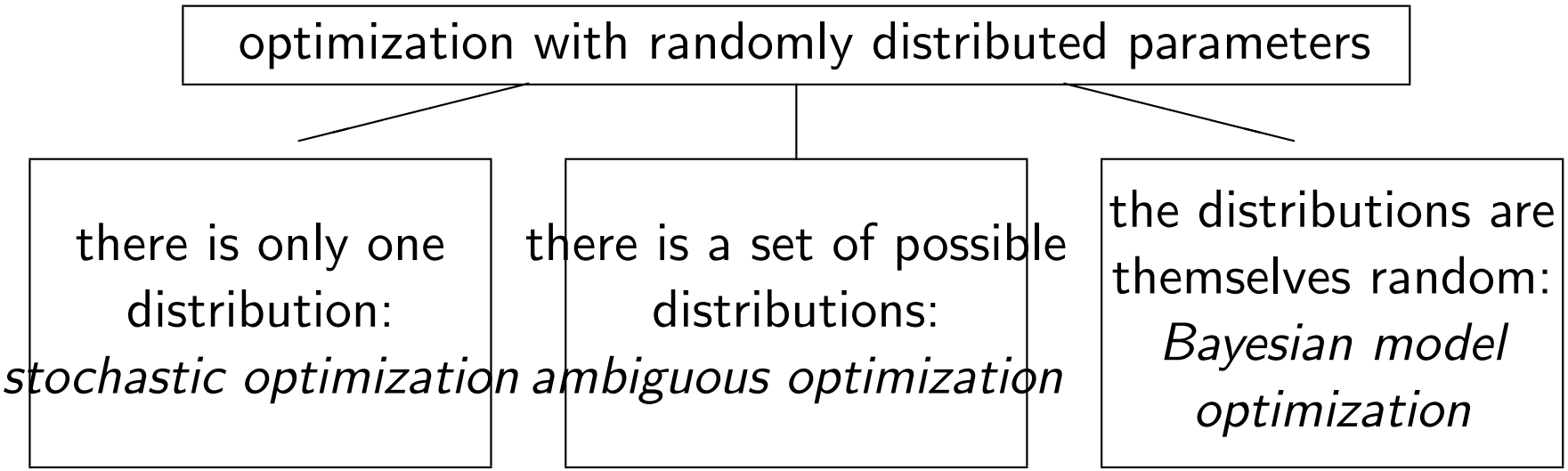


A saddle point

Statistical estimation and confidence sets

- The ambiguity set 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.


parametric optimization
there is only one parameter: *standard deterministic optimization*
there is a set of possible parameters: *robust optimization*
the parameters are randomly distributed: *stochastic optimization*


optimization with randomly distributed parameters
there is only one distribution:
stochastic optimization
there is a set of possible distributions:
ambiguous optimization
the distributions are themselves random:
Bayesian model optimization

Dealing with uncertainty

- Deterministic optimization:

$$\max\{F(x) : F_i(x) \geq 0\}$$

F is the objective,
F_i are the constraint functions.

- Robust optimization (maximin approach):
A set Ξ of possible parameters ξ is given. Typically Ξ is chosen such that

$$P\{\xi \in \Xi\} = 1 - \alpha.$$

$$\max\{\min\{F(x,\xi) : \xi \in \Xi\} : F_i(x,\xi) \geq 0; \xi \in \Xi\}$$

Robust optimization produces very conservative decisions.

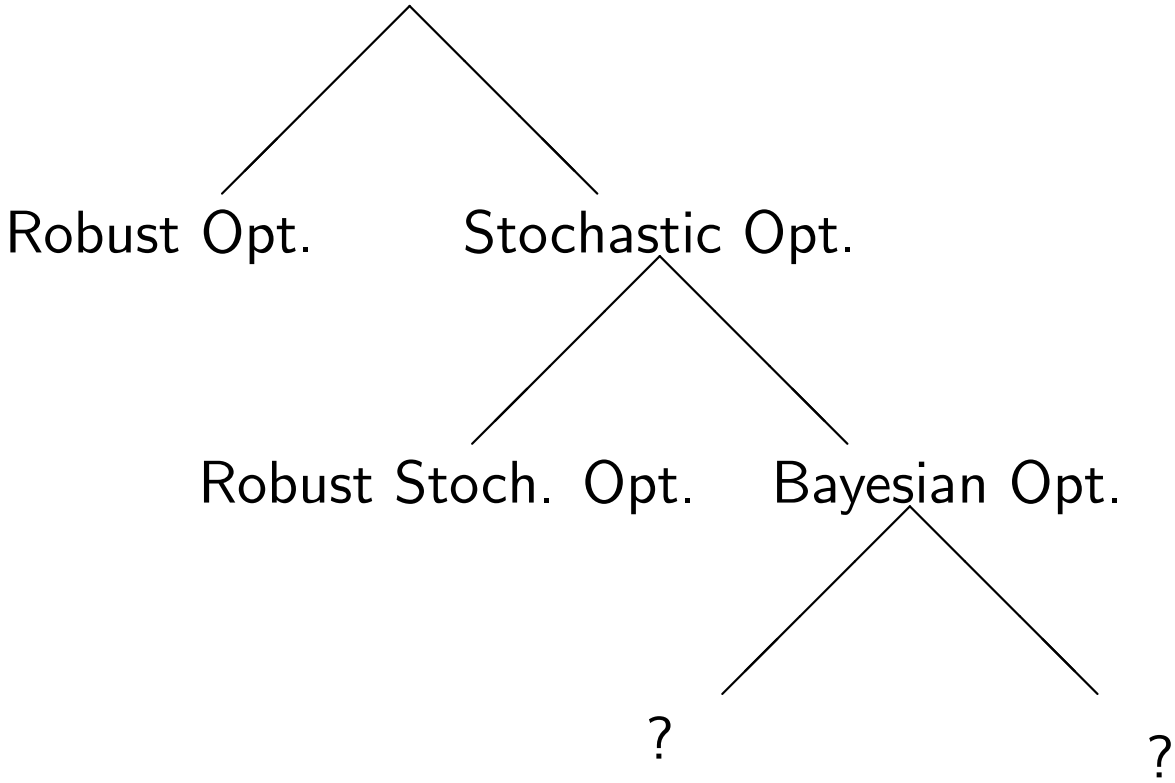
- Stochastic optimization:
  $(\Xi, \mathcal{U}, P)$ is a probability space, typical element is called ξ.

$$\max\{\mathbb{E}_P[F(x,\xi)]; \mathbb{E}_P[F_i(x,\xi)] \geq 0\}$$

or more generally

$$\max\{\mathcal{U}_P[F(x,\xi)]; \mathcal{U}_P^{(i)}[F_i(x,\xi)] \geq 0\}$$

where $\mathcal{U}_P^{(\cdot)}$ are *acceptability functionals*.


Robust Opt.
Stochastic Opt.
Robust Stoch. Opt.
Bayesian Opt.
?
?

Ambiguity models

Shapiro and Kleywegt (2002) define a set of probability measures μsuch that

$$\mathcal{P} = \left\{ \mu : \mu = \sum_{i=1}^{n} \lambda_i \mu_i, \sum_{i=1}^{n} \lambda_i = 1, \lambda_i \geq 0 \right\}.$$

Shapiro and Ahmed (2004) define the ambiguity set as

$$\begin{aligned} \mathcal{P} \quad = \quad & \left\{ \mu \text{ is a prob. measure s.t. } \mu_1 \prec \mu \prec \mu_2, \int \phi_i \, d\mu = b_i; i = 1, \ldots k; \right. \\ & \left. \int \psi \, d\mu \leq c_i; i = 1, \ldots \ell \right\} \end{aligned}$$

where $\mu_1 \prec \mu_2$ means that $\mu_1(A) \leq \mu_2(A)$ for all measurable sets A. In order to allow μ to be probability measures, $\mu_1$must be a positive measure with total mass smaller than 1 and μ_2 has mass larger than 1.

Calafiore (2007) uses the Kullback-Leibler divergence to define

$$\mathcal{P}_\epsilon = \left\{ (p_1, \ldots, p_n) : \sum_{i=1}^{n} p_i \log \frac{p_i}{\hat{p}_i} \leq \epsilon \right\}.$$

Thiele (2007) considers the set

$$\mathcal{P} = \{ (p_1, \ldots, p_n) : |p_i - \hat{p}_i| \leq \epsilon_i \}.$$

Delage and Ye (2010) consider the following ambiguity set

$$\begin{aligned} \mathcal{P} \quad = \quad & \left\{ \mu : \mu(S) = 1, (\int x \, d\mu(x) - c)^\top \Sigma_0^{-1} (\int x \, d\mu(x) - c) \leq \gamma_1, \right. \\ & \left. \int (x - c)(x - c)^\top d\mu(x) \preceq \gamma_2 \Sigma_0 \right\} \end{aligned}$$

Here $\Sigma_1 \preceq \Sigma_2$ means that $\Sigma_2 - \Sigma_1$is a positive definite matrix, i.e. the set is defined by a conical constraint.

Edirishinge (2011) considers ambiguity sets, which are defined by a finite number of generalized moment equalities

$$\mathcal{P} = \left\{ \mu : \int f_i \, d\mu = c_i, i = 1, \ldots, n \right\}.$$

Typically, the requirement is that all elements in the ambiguity set coincide with the baseline probability $\hat{\mu}$ w.r.t. the first n moments. Wozabal and Pflug (2010) use for the first time ambiguity sets, which are balls with respect to the transportation distance.

Wasserstein distance

In order to measure the distance of two scenario distributions we use the transportation distance (Kantorovich distance, Wasserstein distance, earth mover distance) between random distributions on $\mathbb{R}^m = (\Omega, \mathrm{d})$ where d is a distance on $\mathbb{R}^m$.
Wasserstein distance of order r:

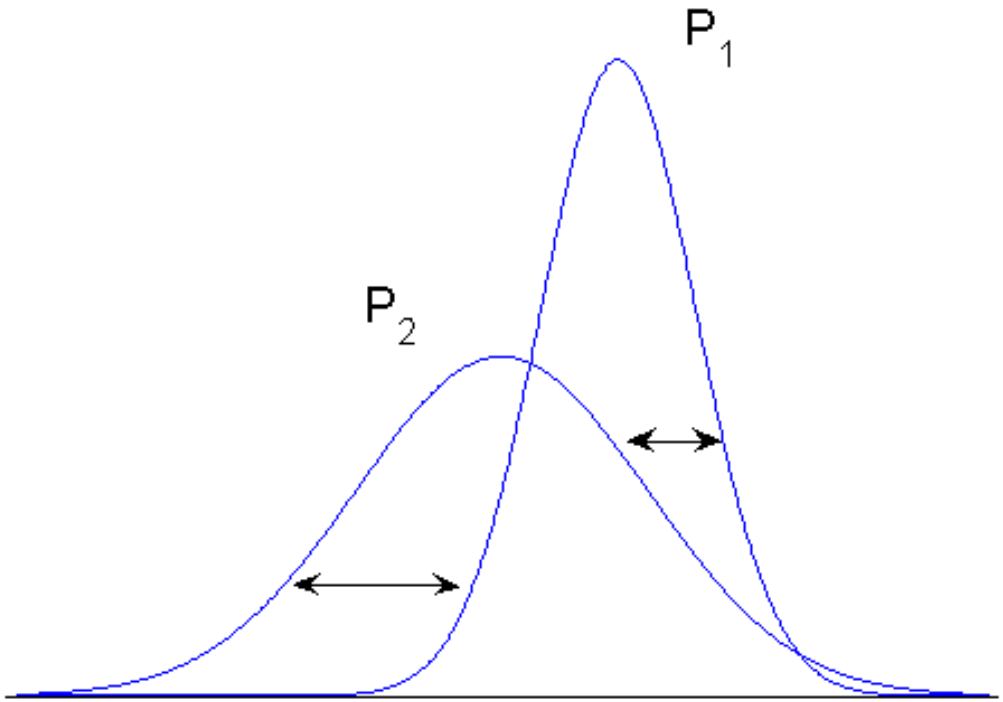
$$\mathsf{dl}_r\left(\mathbb{P}_1, \mathbb{P}_2; d\right) := \left(\inf_{\pi}\left\{\int_{\Omega\times\Omega} d\left(\omega_1, \omega_2\right)^r \pi\left[\mathrm{d}\omega_1, \mathrm{d}\omega_2\right]\right\}\right)^{\frac{1}{r}},$$

where the infimum is taken over all (bivariate) probability measures π on $\Omega \times \Omega$ which have respective marginals, thats

$$\pi\left[A \times \Omega\right] = \mathbb{P}_1\left[A\right] \text{ and } \pi\left[\Omega \times B\right] = \mathbb{P}_2\left[B\right]$$

for all measurable sets $A \subseteq \Omega$ and $B \subseteq \Omega$.
We shall call such a measure π a *transportation plan*.

Illustration of the Wasserstein distance



Why Wasserstein balls?

- If the probability space Ω is finite, $\Omega = \{\omega^{(1)}, \dots, \omega^{(S)}\}$ and the scenario values are kept fixed, then the ambiguity set is a polyhedral set

$$\begin{aligned}\{P : \mathrm{d}_r(P, \hat{P}) \le K\} \quad = \quad & \{P = (p_1, \dots, p_S) : p_j = \sum_i \pi_{ij}; \\ & \sum_j \pi_{ij} = \hat{p}_i; \pi_{ij} \ge 0; \\ & \sum_{i,j} \|x^{(i)} - x^{(j)}\|^r \pi_{ij} \le K^r\}.\end{aligned}$$

- If also the scenario values may vary, the problem is more involved, but can be attacked by DC-algorithms (D. Wozabal)
- The Wasserstein distance metricizes the weak topology on uniformly r-integrable sets of probability measures. In particular, $\mathrm{d}_r(P, \hat{P}_n) \to 0$ for the empirical measure $\hat{P}_n$.

Statistical estimation and confidence sets

- The distance used in the ambiguity should not be too coarse to avoid too large ambiguity sets.
- It should also be not too fine, otherwise no statistical confidence set can be constructed without further assumptions (e.g. for the variational distance)

For the empirical probability measure $\hat{P}_n$, Dudley (1969) has shown that

$$\mathbb{E}_P[d_r(P, \hat{P}_n)] \leq C_r n^{-1/m}$$

for some constant C . By Markov's inequality

$$P\{d_1(P, \hat{P}_n) \geq \epsilon\} \leq \mathbb{E}[d(P, \hat{P}_n)]/\epsilon \leq n^{-1/M} C/\epsilon.$$

Under smoothness conditions on P, the confidence sets may improved (Kersting (1978)).

The ambiguous 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 functional

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

Stress testing cannot replace ambiguity modelling

We solve the portfolio problem for the baseline probability model, but check its performance under new probability models.

Example: 500 historic weekly returns from NYSE

model	composition optimal for	weekly return	acceptabilty $\mathbb{A}$V@R$_{0.1}$
$(\hat{P})$	$(\hat{P})$	0.8%	0.92
(P^-)	$(\hat{P})$	-0.13%	0.90
(P^+)	$(\hat{P})$	1.17%	0.93
(P^-)	(P^-)	-0.07%	0.92
(P^+)	(P^+)	1.9%	0.92

$(\hat{P})$: historic data

(P^-) : 5% higher prob. for bad scenarios

(P^+) : 5% lower prob. for bad scenarios

Equal weights is maximin for large ambiguity

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

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

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.

Solution techniques for the maximin portfolio problem

Let $\mathbb{X} = \{x : x^\top \mathbb{1} = 1, x \geq 0, \mathcal{U}_P(Y_x) \geq q \text{ for all } P \in \mathcal{P}\}$,
then the ambiguity problem reads

$$\max_{x \in \mathbb{X}} \min_{P \in \mathcal{P}} \mathbb{E}_P[Y_x].$$

By continuity and concavity of $\mathcal{U}$, $\mathbb{X}$ is a compact convex set. Moreover, $(P, x) \mapsto \mathbb{E}_P[Y_x]$ is bilinear in P and x and hence convex-concave. Therefore $x^* \in \mathbb{X}$ is a solution of if and only if there is a $Q^* \in \mathcal{P}$ such that (P^*, x^*) is a saddle point, i.e.

$$\mathbb{E}_{P^*}[Y_x] \leq \mathbb{E}_{P^*}[Y_{x^*}] \leq \mathbb{E}_P[Y_{x^*}]$$

for all $(P, x) \in \mathcal{P} \times \mathbb{X}$.
Our problem is semi-infinite. Direct saddle point methods were used by Rockafellar (1976), Nemirovskii and Yudin (1978). We solve it by successive convex programming (SCP).

Successive convex programming (SCP)

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

$$\left\| \begin{array}{l} \text{Maximize (in } x, t) : \ t \\ \text{subject to} \\ t \le \mathbb{E}_P(Y_x) \text{ for all } P \in \mathcal{P}_n \\ \mathcal{U}_P(Y_x) \ge q \ \text{ for all } P \in \mathcal{P}_n \\ x^\top \mathbb{1} = 1; x \ge 0 \end{array} \right.$$

   and call the solution (x_n, t_n).
3. Solve the first inner problem

$$\left\| \begin{array}{l} \text{Minimize (in } P) : \ \mathbb{E}_P(Y_{x_n}) \\ \text{subject to} \\ P \in \mathcal{P} \end{array} \right.$$

   and call the solution $P_n^{(1)}$.

4. Solve the second inner problem

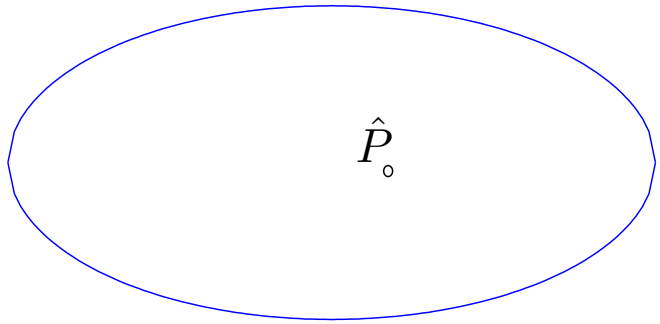
$$\left\| \begin{array}{l} \text{Minimize (in } P\text{)} : \mathcal{U}_P(Y_{x_n}) \\ \text{subject to} \\ P \in \mathcal{P} \end{array} \right.$$

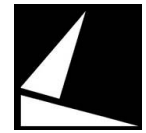
call the solution $P_n^{(2)}$ and let $\mathcal{P}_{n+1} = \mathcal{P}_n \cup \{P_n^{(1)}\} \cup \{P_n^{(2)}\}$.

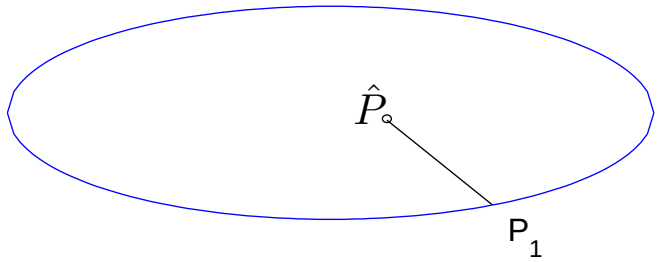
5. If $\mathcal{P}_{n+1} = \mathcal{P}_n$ then stop. Otherwise set $n := n + 1$ and goto 2.

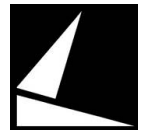
Convergence

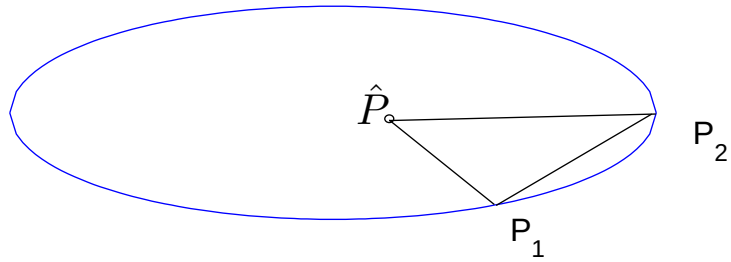
Proposition. Assume that $\mathcal{P}$ is compact and convex and that $(P, x) \mapsto \mathbb{E}_P[Y_x]$ as well as $(P, x) \mapsto \mathcal{U}_P[Y_x]$ are jointly continuous. Then every cluster point of (x_n) is a solution of the maximin problem. If the saddle point is unique, then the algorithm converges to the optimal solution.

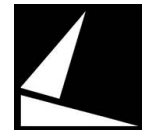

$\hat{P}_\circ$

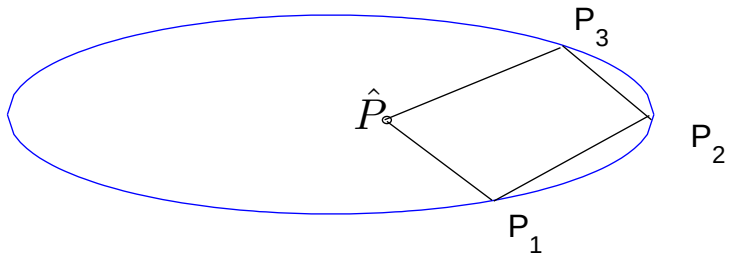


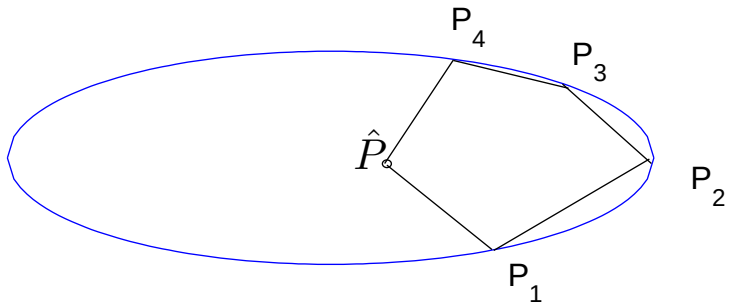

$\hat{P}$
P_1




$\hat{P}$
P 2
P 1




P 3
$\hat{P}$
P 2
P 1


P4
P3
$\hat{P}$
P2
P1

Example

6 assets

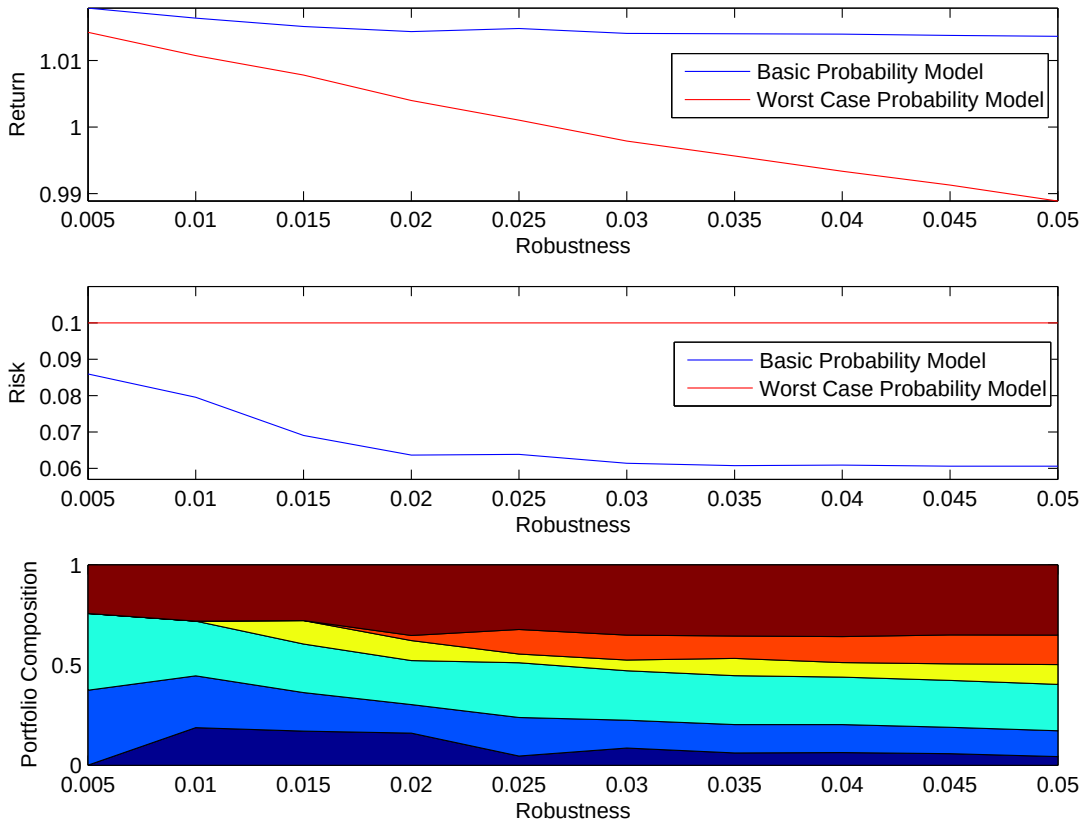
- IBM - International Business Machines Corporation
- PRG - Procter & Gamble Corporation
- ATT - AT&T Corporation
- VER - Verizon Communications Inc
- INT - Intel Corporation
- EXX - Exxon Mobil Corporation

Risk constraint: $\mathbb{A}\mathrm{V@R}_{0,1} \geq 0.9$
Ambiguity set: $\mathcal{P} = \{P : \mathrm{d}_1(P, \hat{P}) \leq \epsilon\}$.
In order to make the ambiguity sets more interpretable, we define a robustness parameter γ as the maximal relative change of the expected returns and relate this to ϵ by

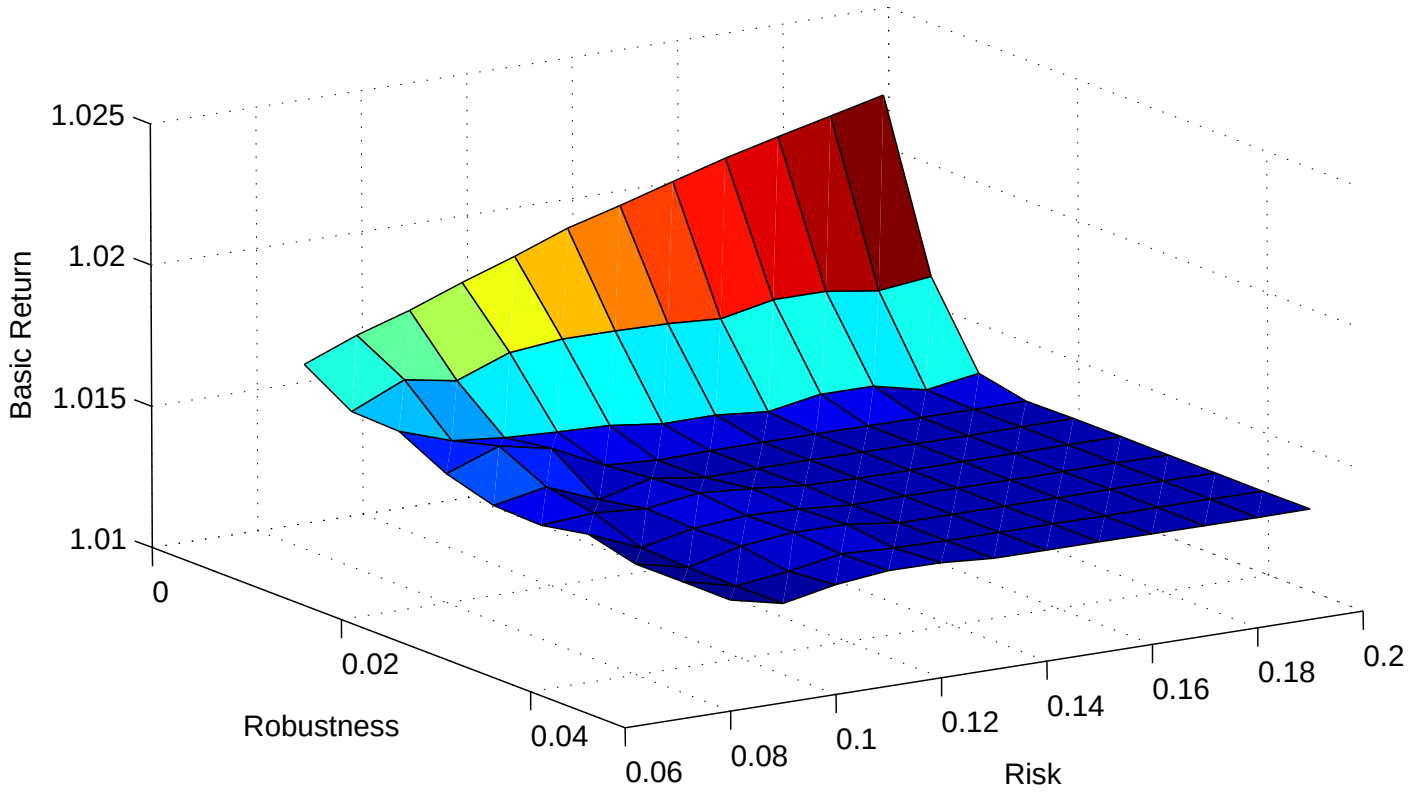
$$\epsilon = \max\{\eta : \sup_{\mathrm{d}_1(P,\hat{P})\leq\eta} \mathbb{E}_P(\xi^{(i)}) \leq (1+\gamma)\mathbb{E}_{\hat{P}}(\xi^{(i)}) : \text{ for all } i\}.$$

γ is used in the figures.

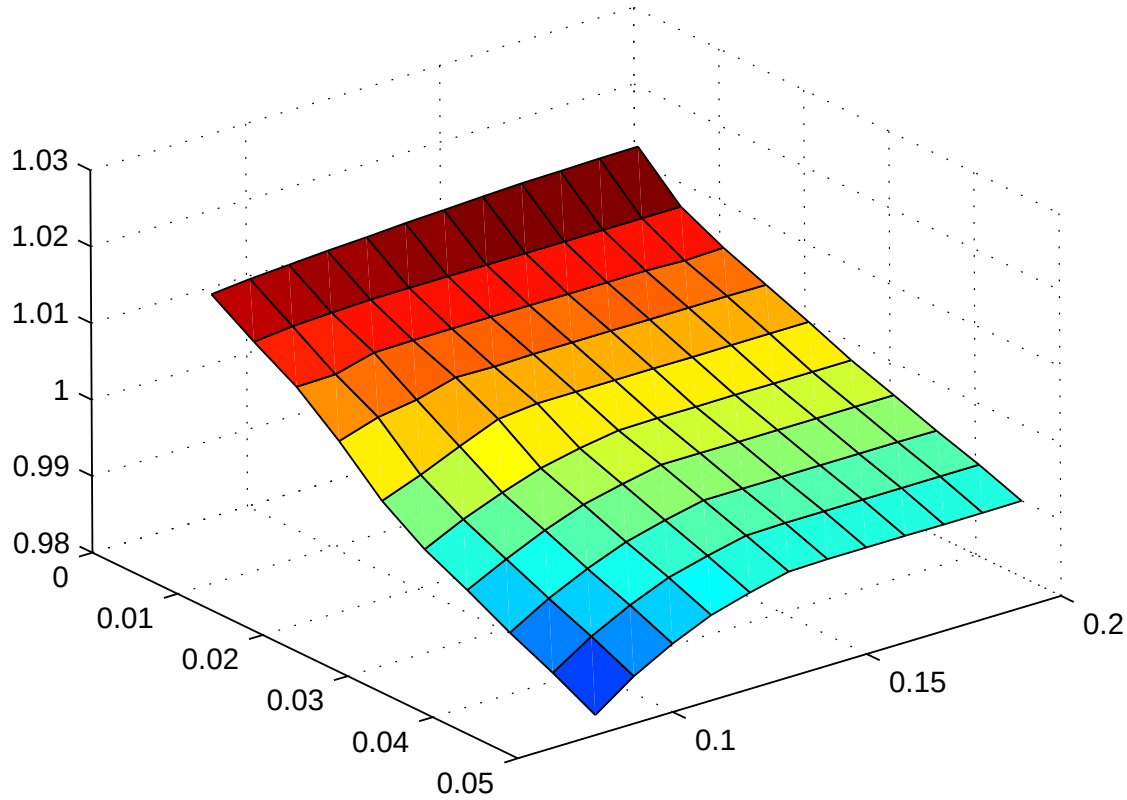
The solution of the Example



Expected returns, risks and portfolio composition. The assets from top to bottom are: EXX, VER, ATT, PRG, INT, IBM.



Efficient frontiers in dependence of the robustness parameter γ.
Risk and return are calculated w.r.t. the basic model P.



Efficient frontiers in dependence of the robustness parameter γ.
Risk and return are calculated w.r.t. the worst case model P^*.

Multistage models

As before, a baseline problem

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

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

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

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

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

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

The nested distance

Recall that the nested distance $\mathrm{dl}_r(\mathbb{P}, \bar{\mathbb{P}})$ is the minimal value of the following optimization program

$$\mathrm{dl}_r(\mathbb{P}, \bar{\mathbb{P}}) = \min \left\{ \left[\int \mathrm{d}_r^r(\xi, \bar{\xi})\, \pi(d\xi, d\bar{\xi}) \right]^{1/r} : \pi \text{ fulfills (1) and (2)} \right\}$$

$$\pi(M \times \bar{\Omega} | \mathcal{F}_t \otimes \bar{\mathcal{F}}_t) = P(M|\mathcal{F}_t) \qquad M \in \mathcal{F}_T \tag{1}$$

$$\pi(\Omega \times N | \mathcal{F}_t \otimes \bar{\mathcal{F}}_t) = \bar{P}(N|\bar{\mathcal{F}}_t) \qquad N \in \bar{\mathcal{F}}_T. \tag{2}$$

To each transportation plan π there correspond transportation subplans in the following manner: If $m \in \mathcal{N}_t, n \in \bar{\mathcal{N}}_t$ and if $k \in m+, \ell \in n+$then $\pi(k, \ell | m, n)$ is defined as

$$\pi(k, \ell | m, n) = \frac{\sum_{i \succ k, j \succ \ell} \pi(i, j)}{\sum_{i \succ m, j \succ n} \pi(i, j)}.$$

Transportation kernels

If π is a transportation plan, then the corresponding subkernel for transporting $\hat{P}$ to P is

$$K(j|i) = \frac{\pi(i,j)}{\sum_j \pi(i,j)} = \frac{\pi(i,j)}{\hat{P}(i)}$$

In a similar manner, transportation subkernels can be defined. If $\pi(k,\ell|m,n)$ is a transportation subplan, then the corresponding subkernel is

$$K(\ell|k;m,n) := \frac{\pi(k,\ell|m,n)}{\sum_\ell \pi(k,l|m,n)} = \frac{\pi(k,\ell|m,n)}{\hat{P}(k)}.$$

In order to indicate the stage t, we may explitely write $K_t(\cdot|k,m,n)$ if $m \in \mathcal{N}_t$, $n \in \bar{\mathcal{N}}_t$.

Notice that the subkernels satisfy

$$K_t(\ell \,|\, k; m, n) \geq 0, \sum_{\ell \in n+} K_t(\ell \,|\, i; m, n) = 1$$

and can be interpreted as Markov transition operators.
The transportation plan π satisfies the "filtration constraints", iff

$$\begin{aligned} K(j|i) &= K_1 \circ \cdots \circ K_{T-1}(j|i) \\ &= K_1(j_1) \,|\, i_1; 1, 1) \cdot \cdots \cdot K(j_{T-1} \,|\, i_{T-1}; i_{T-2}, j_{T-2}) \times \\ &\quad \times K(j \,|\, i; i_{T-1}, j_{T-1}). \end{aligned}$$

We write $P = K \circ \hat{P} = K_1 \circ \cdots \circ K_{T-1}\hat{P}$ for the concatenation of the subkernels. The problem of probability ambiguity can now be written as

$$\max_x \min_K \left\{ \mathcal{U}_{K\circ\hat{P}}[Q(x,\xi)] \,:\, K = K_1 \circ \cdots \circ K_{T-1}; \right.$$
$$\left. \sum_{i,j\in\mathcal{N}_T} \mathrm{d}_\xi(i,j)\, K(j|i)\, \hat{P}(i) \le \varepsilon \right\},$$

where $d_\xi(i,j) = \sum_{t=1}^{T} \|\xi(i_t) - \xi(j_t)\|$. Notice that each transportation subkernel is a linear mapping, but the composition is multilinear in $K_1, \ldots K_{T-1}$.

The algorithm

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

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

   and call the solution x_n.
3. Solve the inner problem.

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

   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.

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

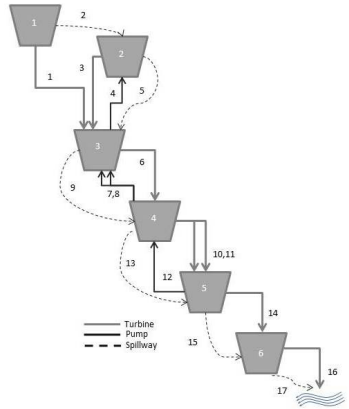
"How much do I loose by implementing the minimax strategy $x^*(\mathcal{P})$ instead of the best strategy for the baseline model, if in fact the baseline model is true?"

- Reward for robust decisions.

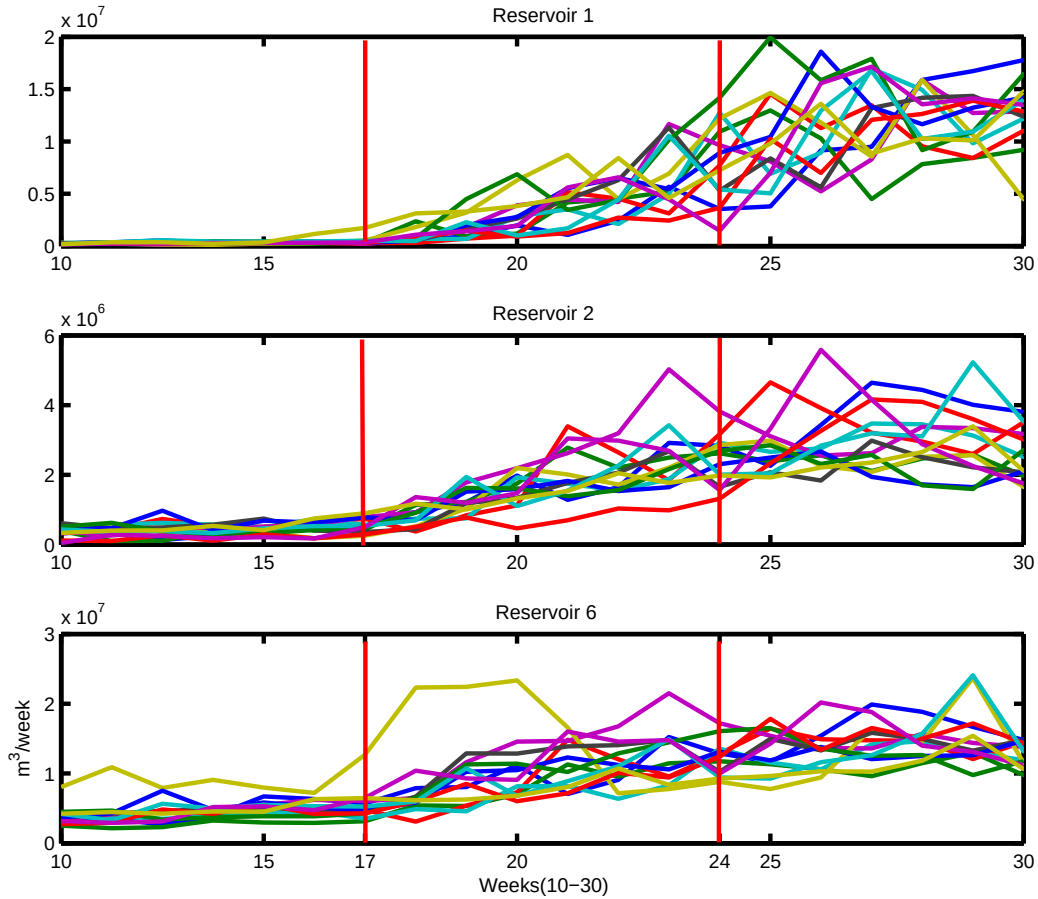
$$\mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathbb{P}),\xi)] - \mathbb{E}_{\mathbb{P}^*}[Q(x^*(\mathcal{P}),\xi)] \geq 0.$$

"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?"

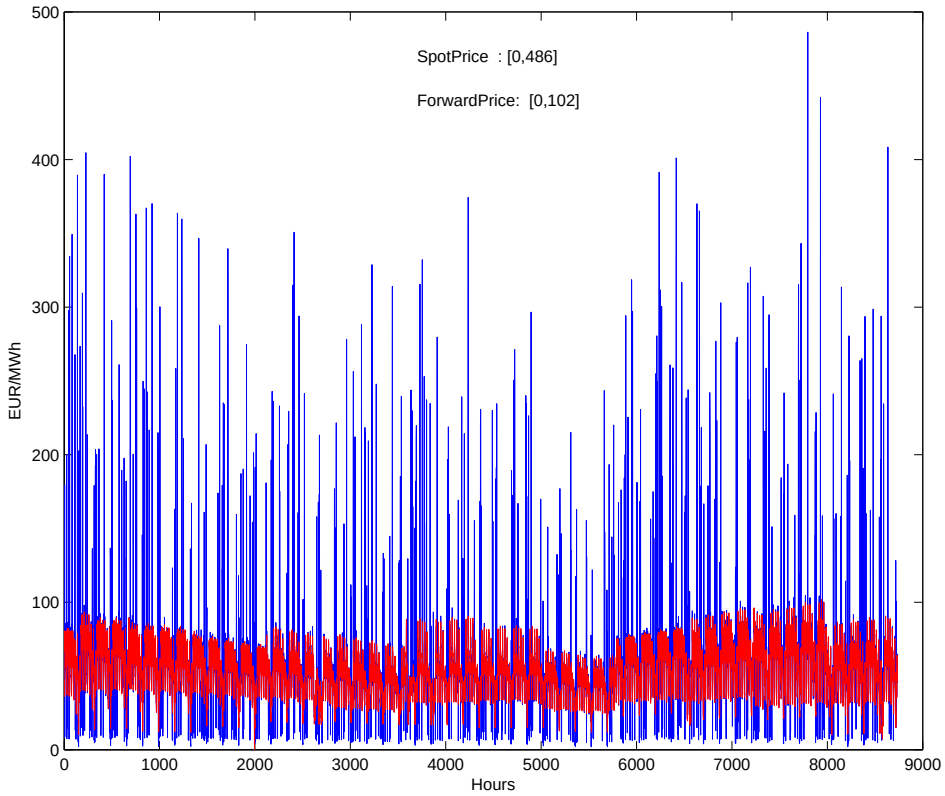
Management of a hydrosystem in the Austrian Alps



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 $SARMA(1,2),(2,2)_5 2$ 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

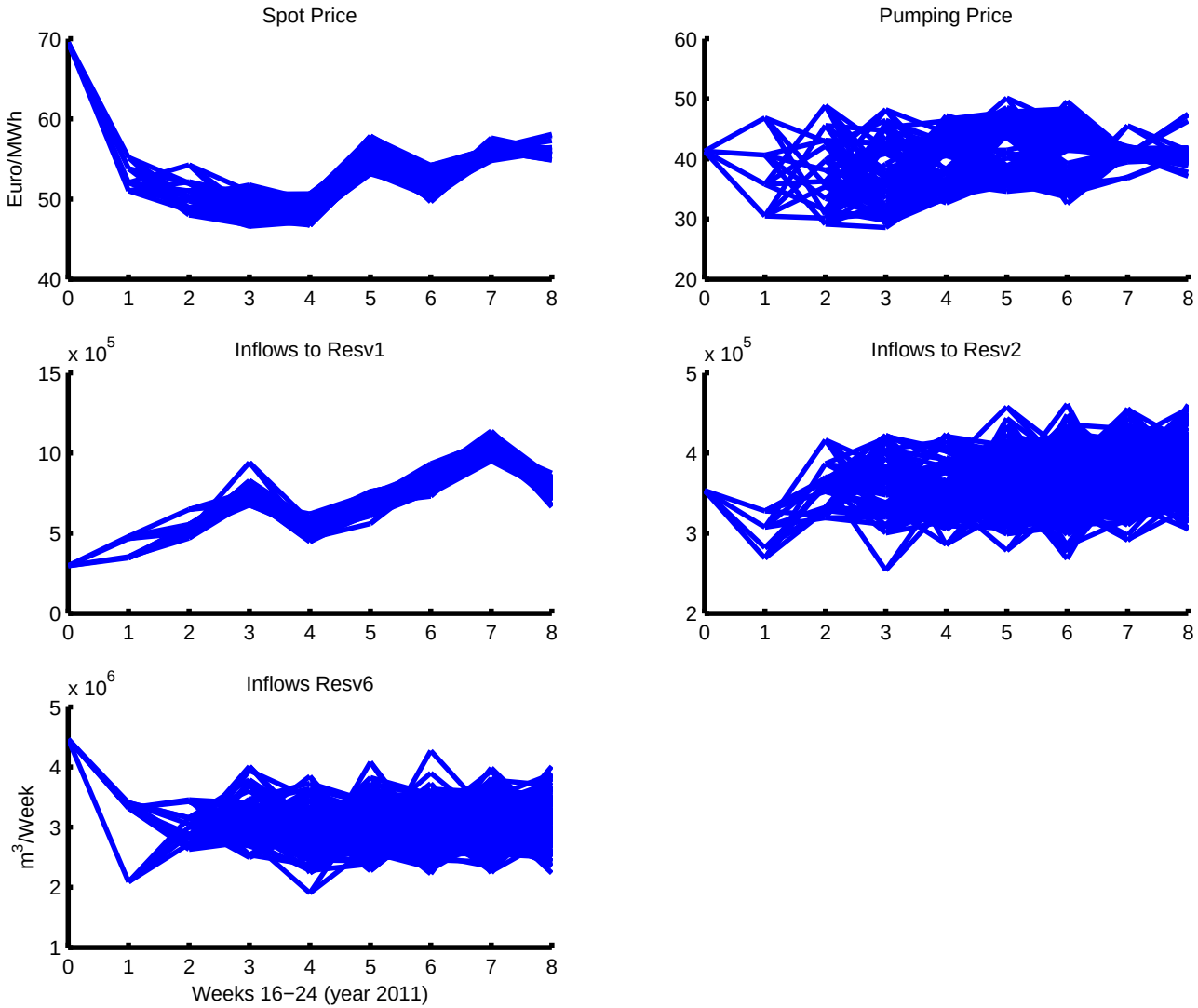
The decision model

maximize
$\lambda\,\mathbb{E}[x_T^c] - (1-\lambda)\mathbb{A}\mathrm{V@R}_{1-\alpha}[-x_T^c]$
subject to
$0 \le x_{t,i}^f \le \overline{x}_i^f,$
$\underline{x}_j^s \le x_{t,j}^s \le \overline{x}_j^s,$
$x_{end,j}^{,s} \le x_{T,j}^s,$
$x_{t,j}^s = x_{t-1,j}^s + \xi_{t,j}^f + \sum_{\{i\in I|P_{max}>0\}} A_{i,j}\cdot x_{t-1,i}^f + \sum_{\{i\in I|P_{max}=0\}} A_{i,j}\cdot x_{t,i}^f,$
$x_{t,i}^e = x_{t-1,i}^f \cdot k^i \cdot \triangle t_{(t-1)},$
$x_t^c = x_{t-1}^c \cdot (1+r)^{\triangle t_{(t-1)}} + \sum_{\{i\in I|k^i>0\}} x_{t-1,i}^e \cdot \xi_t^e + \sum_{\{i\in I|k^i<0\}} x_{t-1,i}^i \cdot \xi_t^p.$

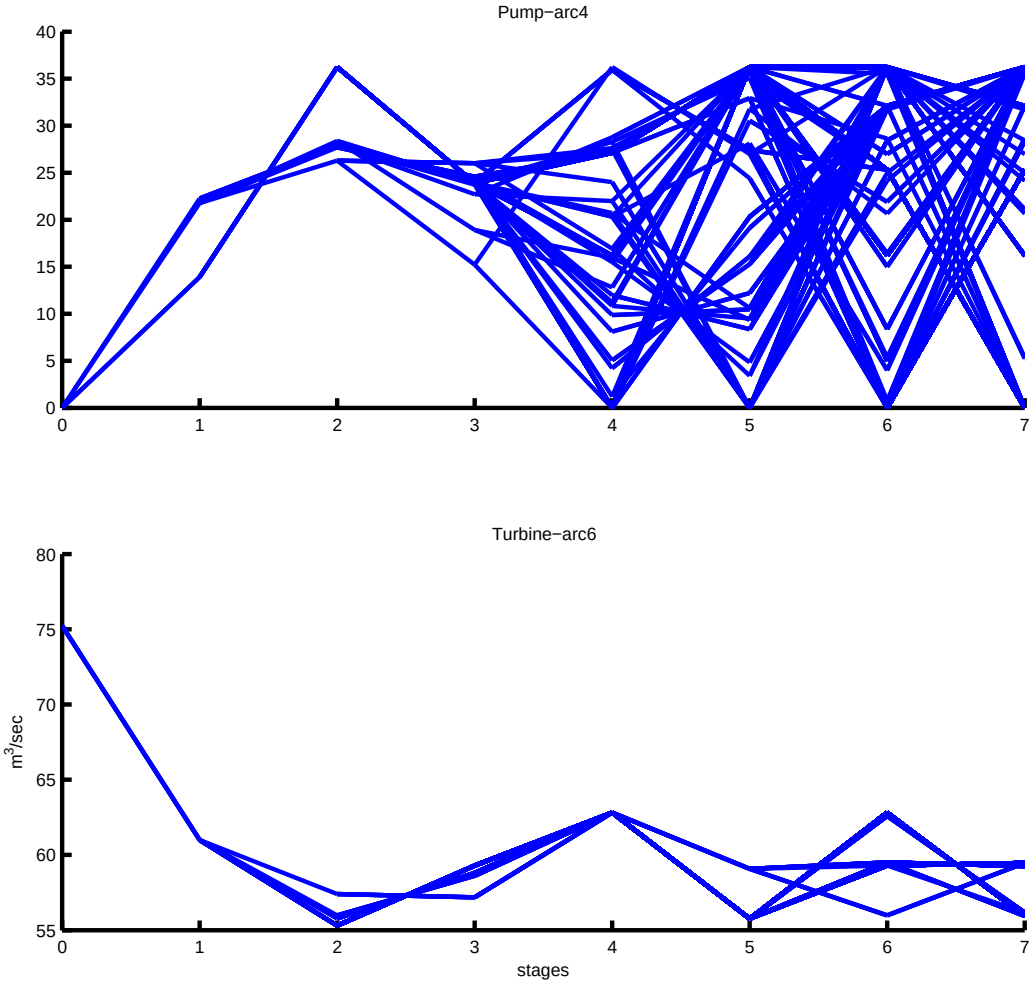
Generating a scenario tree

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.

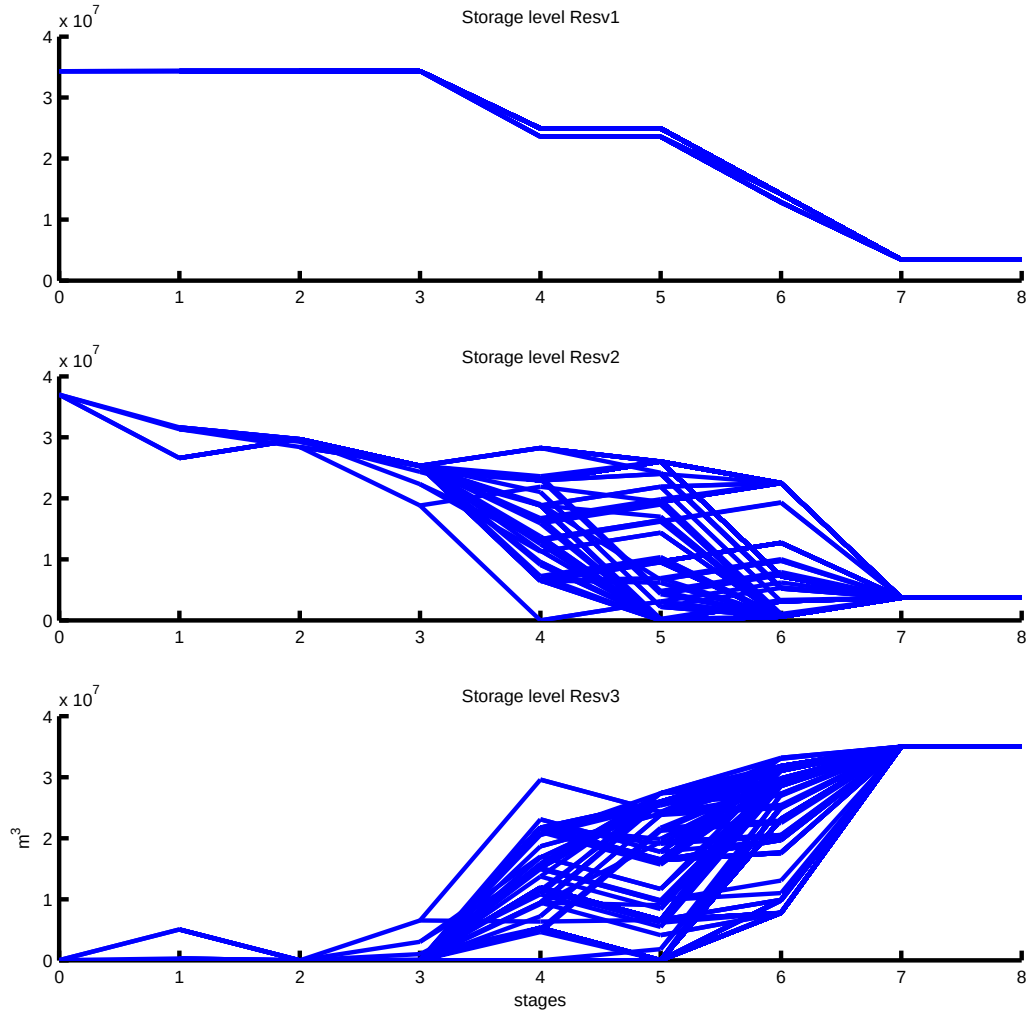
Number of stages	8
Minimal bushiness per stage	2,2,2,1,1,1,1,1
Maximal distance per stage	5,5,5,7,7,7,10,10
Number of scenarios (leaves)	392
Number of nodes 1532	



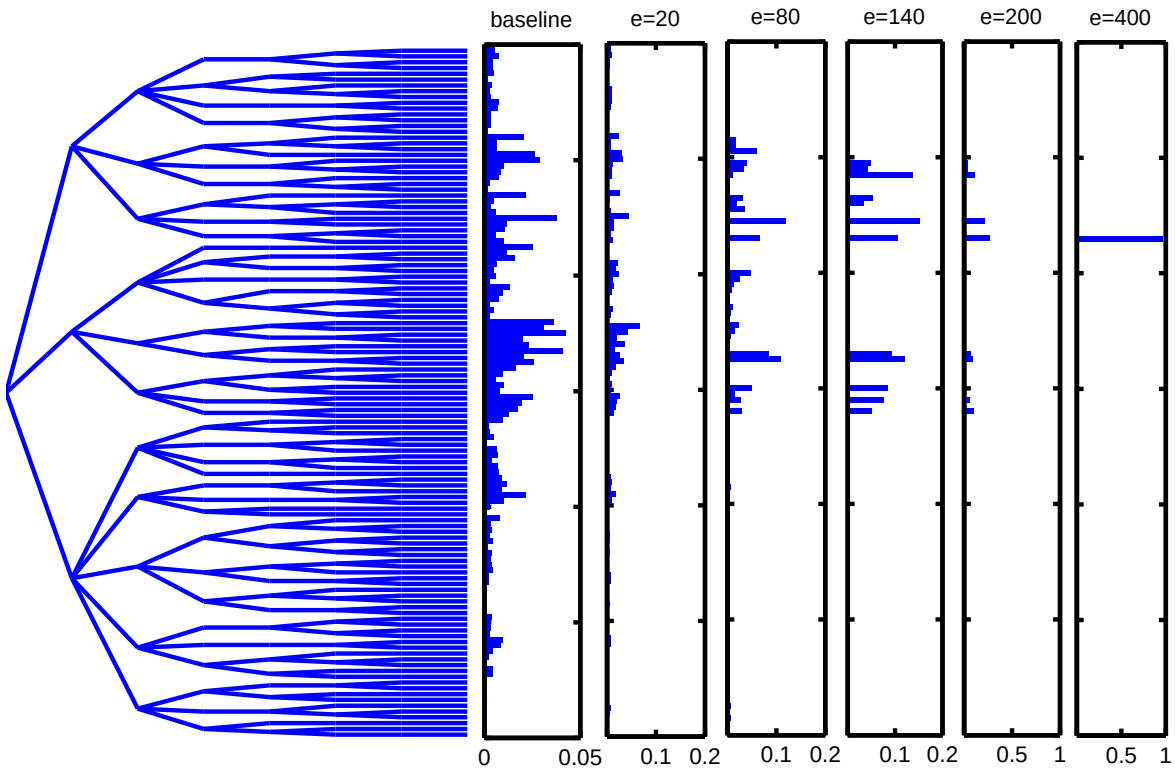
The generated five-dimensional tree



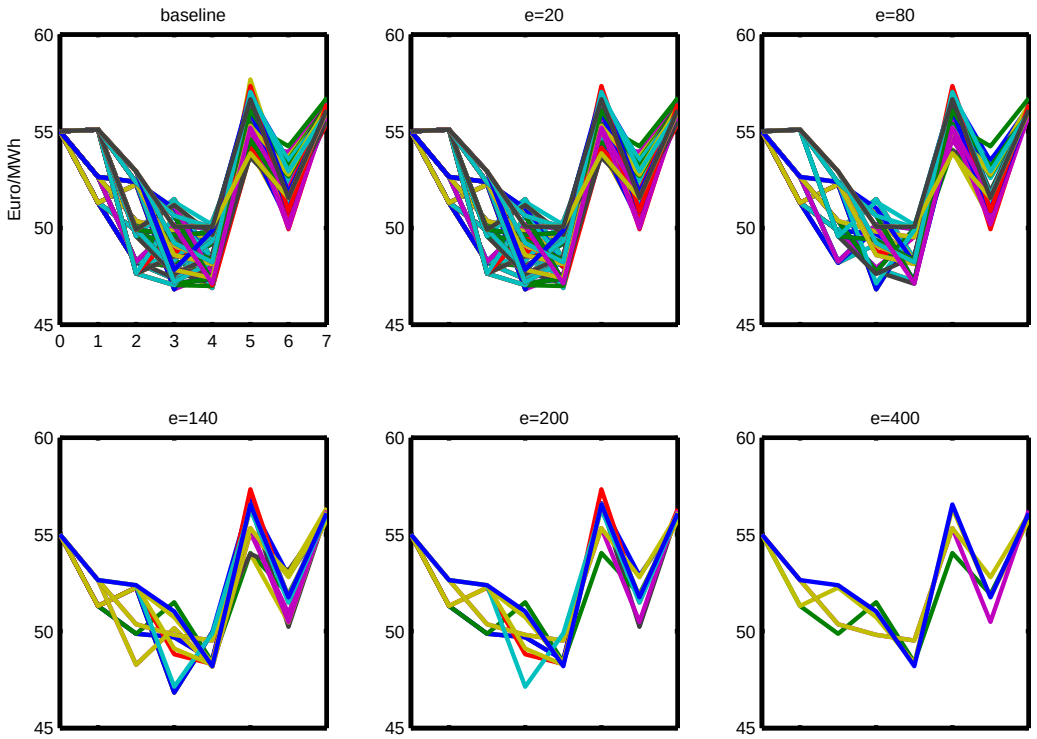
The pumping (top) and turbining (bottom) decisions



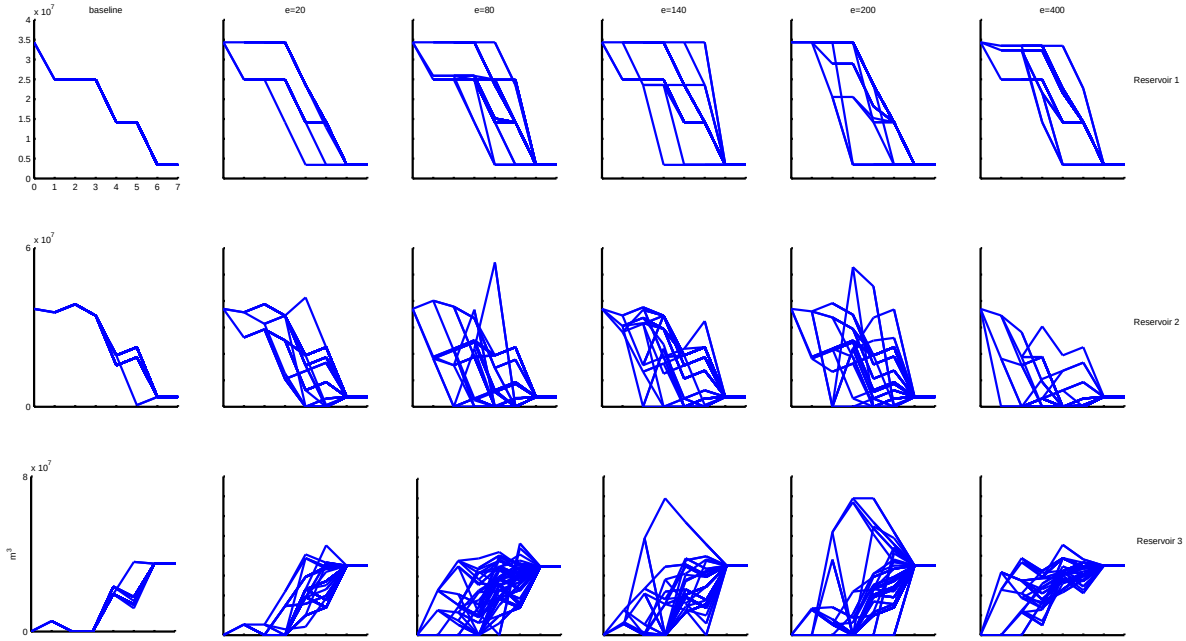
The storage levels



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.

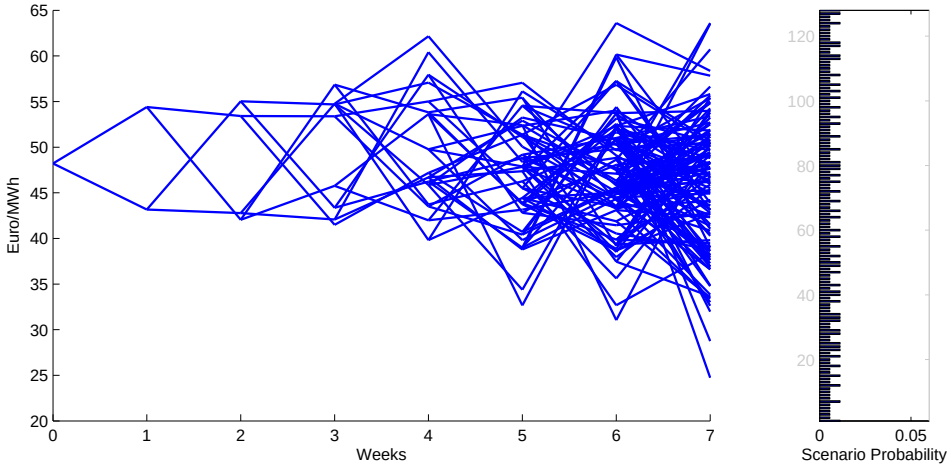


The minimax decisions: They get more complicated with increasing ambiguity radius: Decisions lying on bounds are avoided.

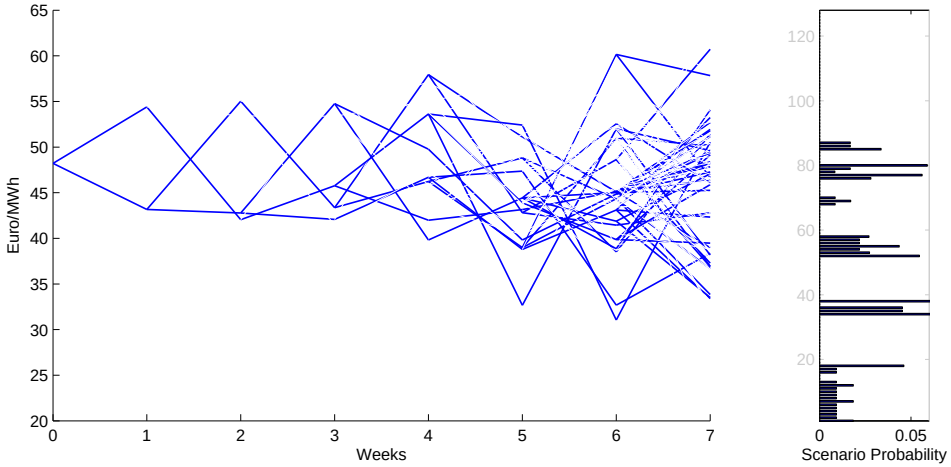
Price of ambiguity: 2.3%.

Reward for robustness: 7.5%.

Worst case tree for a thermal plant optimization



The original spotprice tree



Conclusions

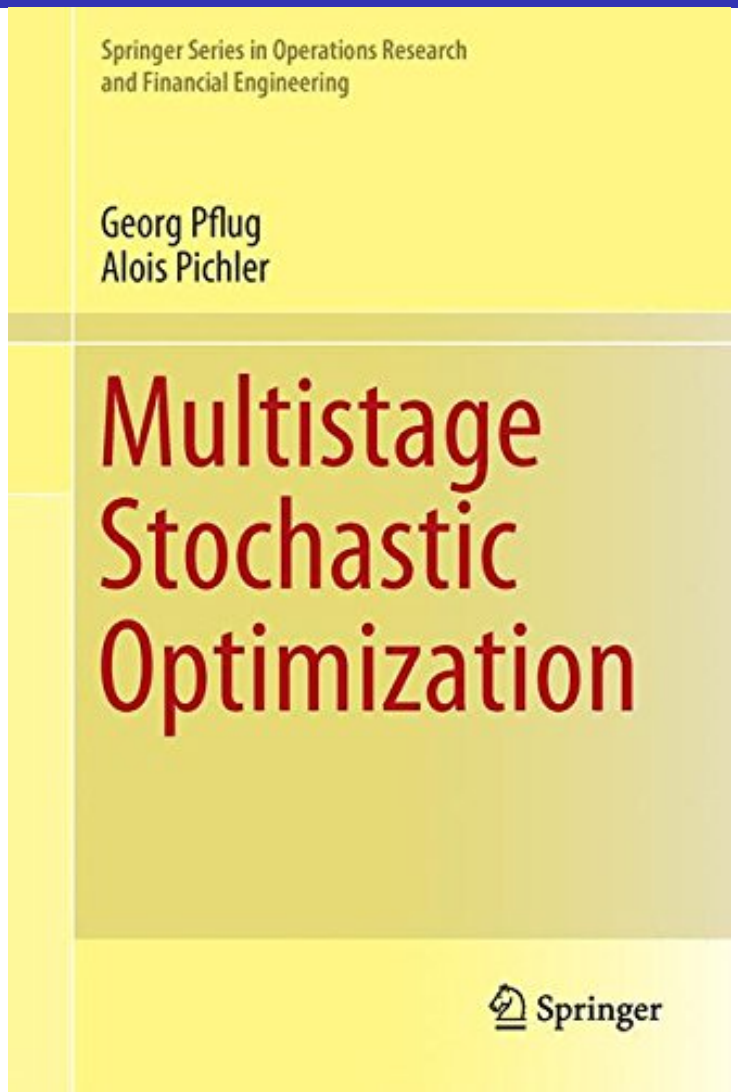
- In order to capture scenario uncertainty (aleatoric uncertainty) and probability ambiguity (epistemic uncertainty) we use a probabilistic maximin approach.
- The ambiguity neighborhood should be chosen in such a way that it corresponds to statistical confidence regions for which bounds for the covering probability are available.
- If the ambiguity radius is increased, then the saddle point changes in the following way:
  - The robust decision strategy becomes more complicated and "diversified"
  - The worst case model gets more simpler
- It turns out that the price to be paid for including ambiguity in the optimization problem is often smaller than the reward one gets for robustifying the solution.
- The same approach may be used for real option or smart grid optimization

Decision making under uncertainty

Georg Ch. Pflug

April 19, 2016

Our Springer book



Robust decisions

Wikipedia: *Robust decision* first used in the late 1990s. It is used to identify decisions made with a process that includes formal consideration of uncertainty. A robust decision is the best possible choice, one found by eliminating all the uncertainty possible within available resources, and then choosing, with known and acceptable levels of satisfaction and risk. (David G. Ullman)
Stochastic optimization methods are optimization methods that generate and use random variables. For stochastic problems, the random variables appear in the formulation of the optimization problem itself, which involve random objective functions or random constraints, for example.

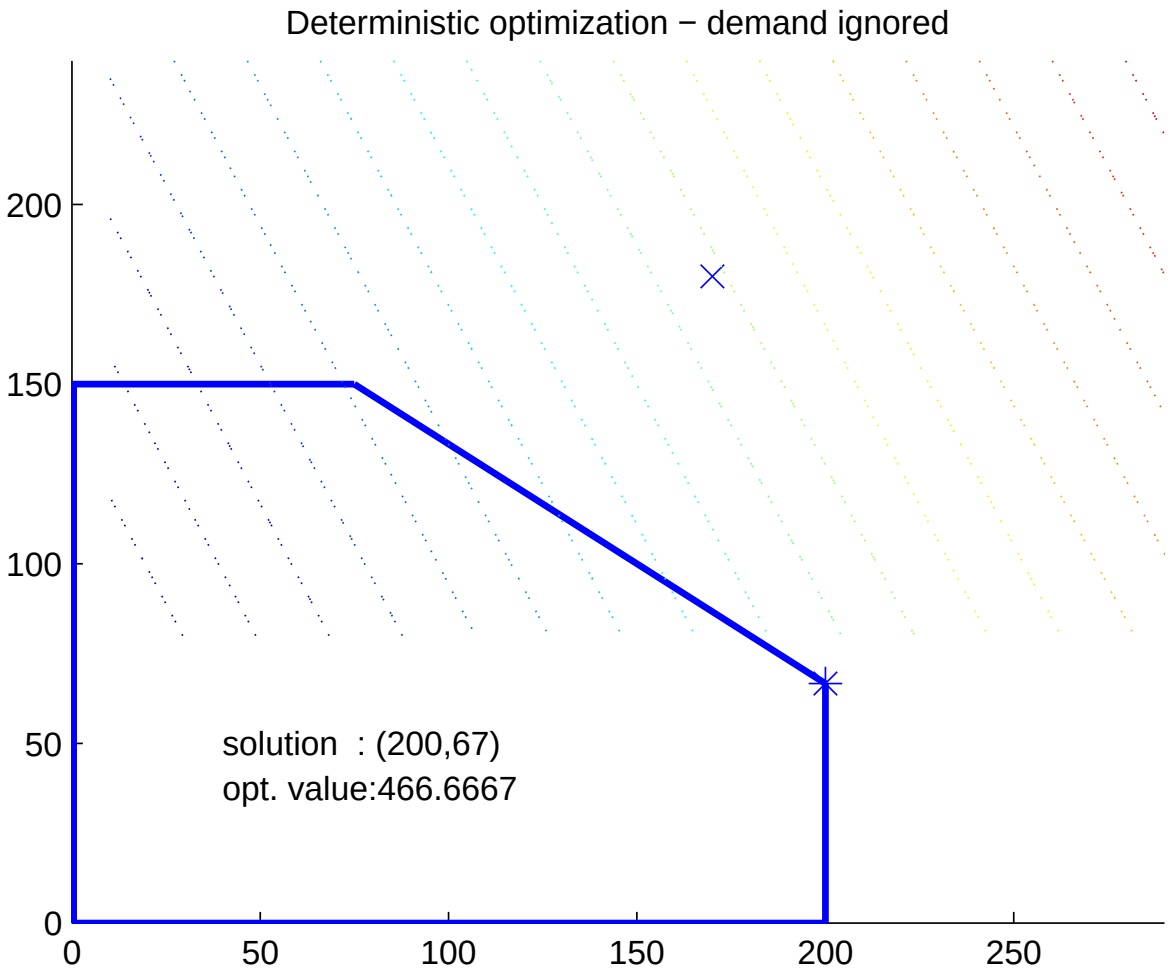
The basic Example

A producer has to determine his production plan for two types of goods. Let

x_1, x_2	the production quantity (decision variable)
$x_1 \leq m_1, x_2 \leq m_2$	individual capacity constraints
$a_1 x_1 + a_2 x_2 \leq b$	joint capacity constraint
c_1, c_2	production cost per unit
s_1, s_2	selling price per unit

Objective: $\max\{(s_1 - c_1)x_1 + (s_2 - c_2)x_2 : x \in \mathbb{X}\}$

with $\mathbb{X} = \{(x_1, x_2) : x_1 \leq m_1, x_2 \leq m_2, a_1 x_1 + a_2 x_2 \leq b\}$ being the feasible set.

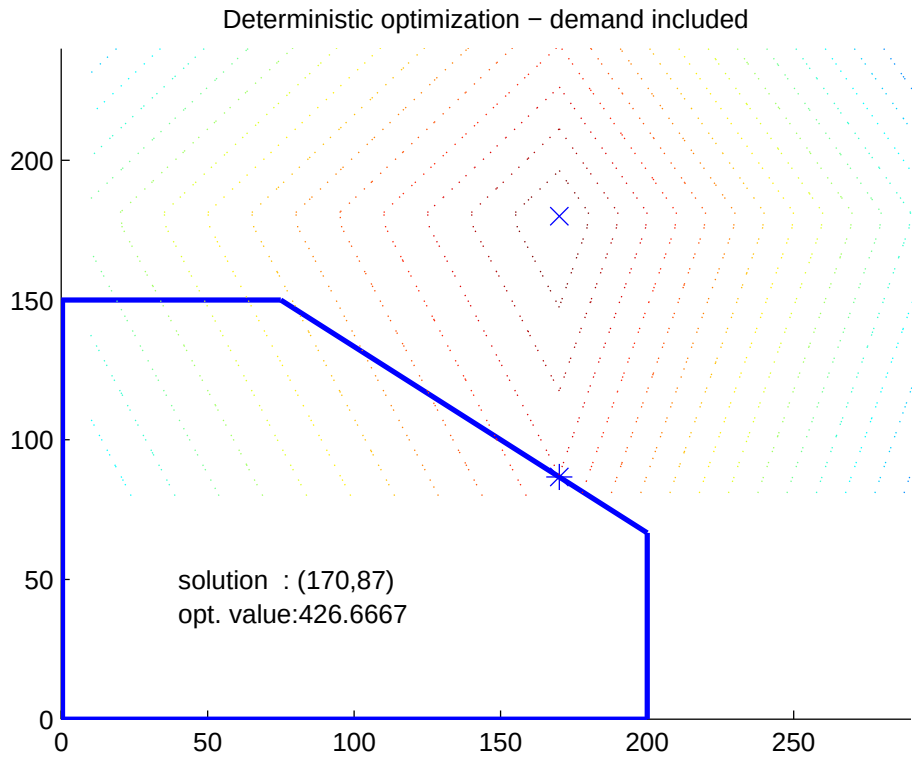

Deterministic optimization − demand ignored
200
150
100
50
0
0
50
100
150
200
250
solution : (200,67)
opt. value:466.6667

Adding a demand value

μ_1, μ_2 the deterministic demands

Objective:

$$\max\{\min(x_1, \mu_1)s_1 + \min(x_2, \mu_2)s_2 - c_1x_1 - c_2x_2 : x \in \mathbb{X}\}$$

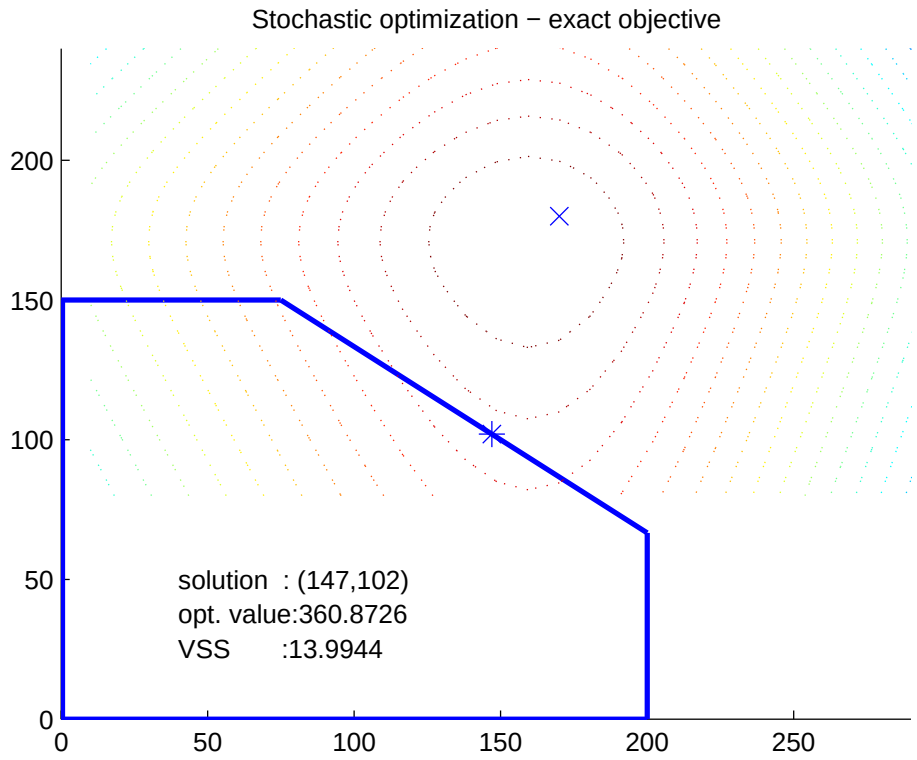


A stochastic demand

$\xi_1 \sim N(\mu_1, \sigma_1^2), \xi_2 = N(\mu_2, \sigma_2^2)$ the random demands

Objective:

$$\max\{\mathbb{E}[\min(x_1, \xi_1)s_1 + \min(x_2, \xi_2)s_2 - c_1 x_1 - c_2 x_2] : x \in \mathbb{X}\}$$



The value of the stochastic solution (VSS)

Let (x_1^+, x_2^+) be the solution of the deterministic problem (where all random variables are set equal to their expectation - in our case the demand limits). The VSS value is

$$\begin{aligned} VSS &= \max\{\mathbb{E}[\min(x_1,\xi_1)s_1 + \min(x_2,\xi_2)s_2 - c_1x_1 - c_2x_2] : x \in \mathbb{X}\} \\ &- \mathbb{E}[\min(x_1^+,\xi_1)s_1 + \min(x_2^+,\xi_2)s_2 - c_1x_1^+ - c_2x_2^+] \end{aligned}$$

The clairvoyant's solution

The decision of the stochastic decision problem has to be made before the demand is observed (here-and-now). A clairvoyant would be able to make the decision after the observation of the demand (wait-and-see), i.e. he solves the following problem:

$$\begin{aligned} & (x_1(\xi), x_2(\xi) \\ = \quad & \operatorname{argmax}_{x_1,x_2}\{\min(x_1,\xi_1)s_1 + \min(x_2,\xi_2)s_2 - c_1x_1 - c_2x_2 : x \in \mathbb{X}\} \end{aligned}$$

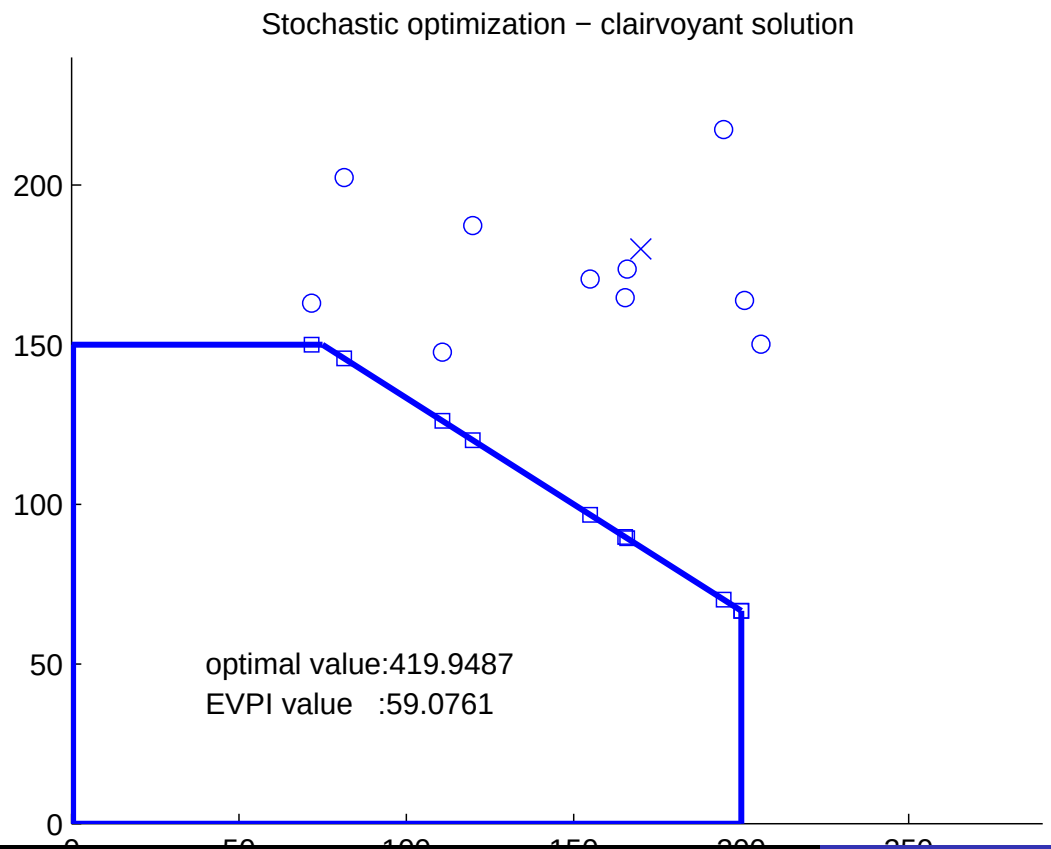
with $\xi = (\xi_1, \xi_2)$.
The clairvoyant's optimal solution is

$$\mathbb{E}[\min(x_1(\xi),\xi_1)s_1 + \min(x_2(\xi),\xi_2)s_2 - c_1x_1(\xi) - c_2x_2(\xi)]$$

The expected value of perfect information (EVPI)

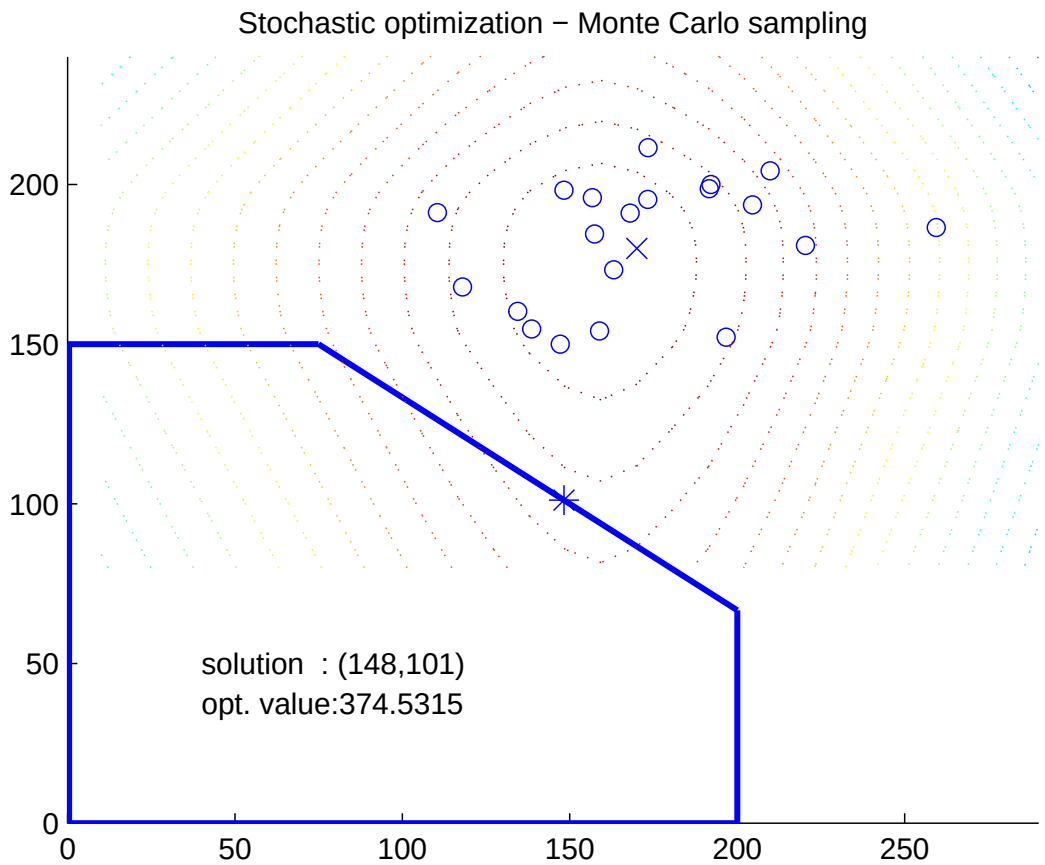
The EVPI value is:

$$\begin{aligned} EVPI &= \text{optimal value of the clairvoyant's problem} \\ &- \text{optimal value of the here-and-now problem} \end{aligned}$$

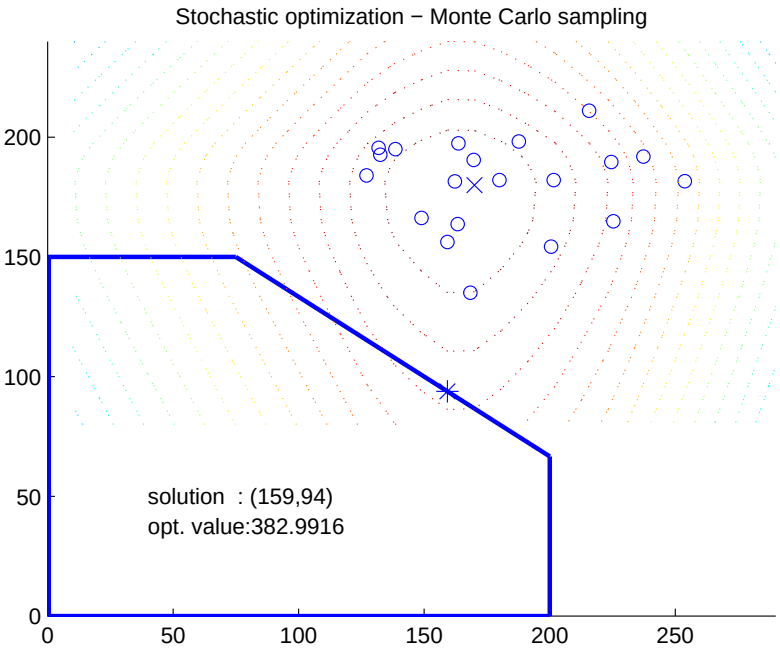


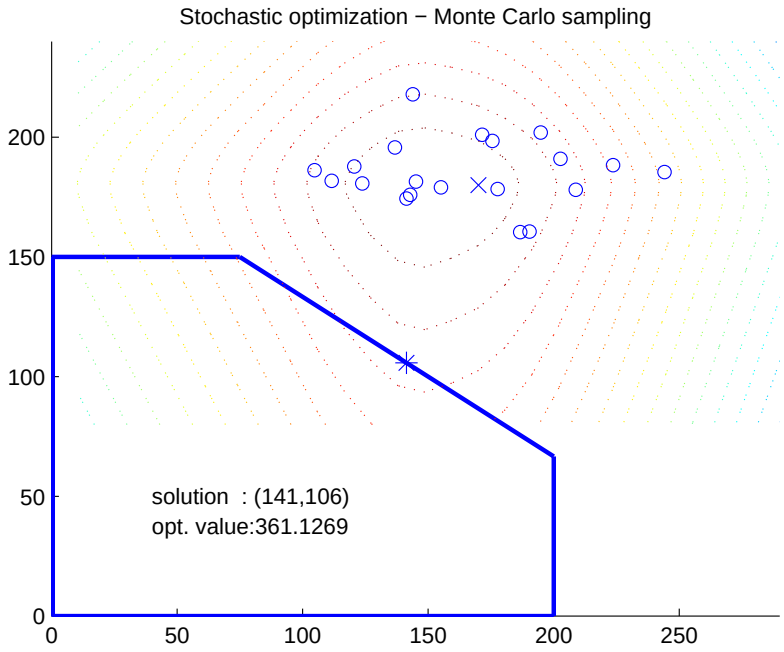
Approximation by MC sampling

Objective: $\max\{\frac{1}{n}\sum_{i=1}^{n}[\min(x_1,\xi_{1,i})s_1+\min(x_2,\xi_{2,i})s_2] : x\in\mathbb{X}\}$



MC approximations give different solution in different runs

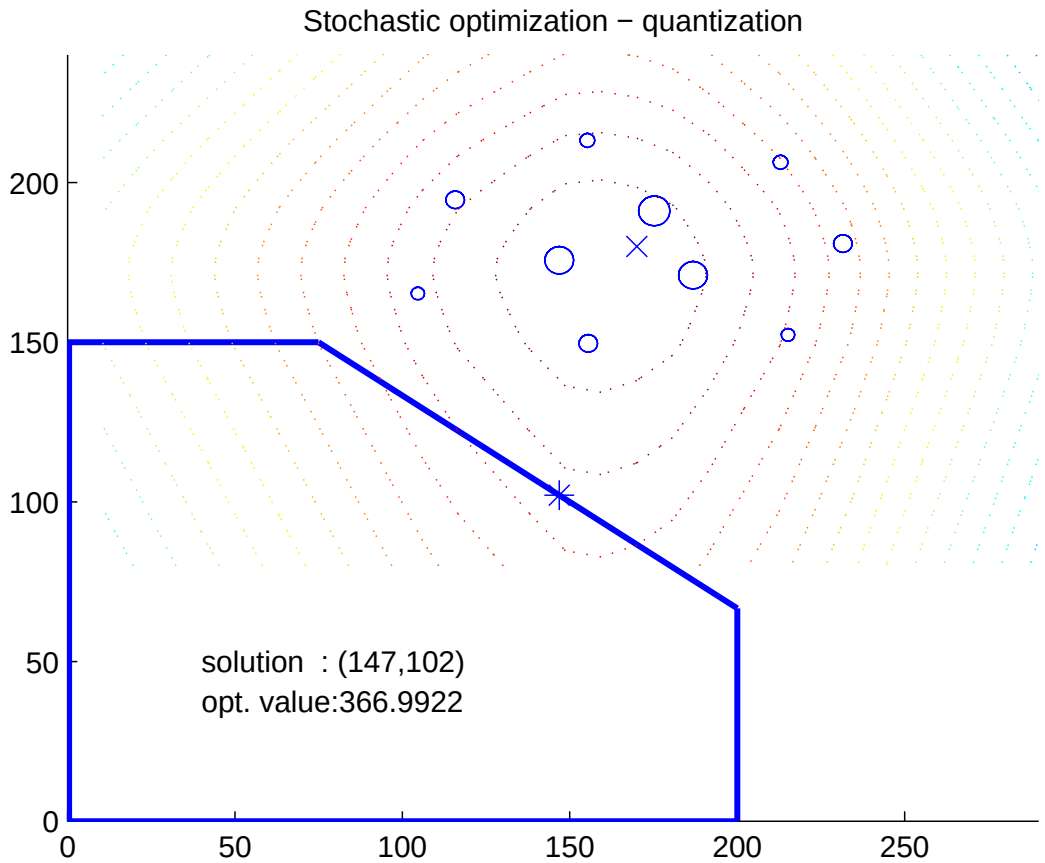




Approximation by quantization

Instead of sampling, one may use optimally placed weighted points, which are chosen in order to minimize a distance between them and the probability distribution of the problem.
To calculate optimal points is a nonlinear nonconvex optimization problem of its own. For the normal distributions, optimal points have been calculated by Gilles Pagès and published on his web-pages (the Pagès-pages).

Objective: $\max\{\sum_{i=1}^{n} p_i[\min(x_1, \xi_{1,i})s_1 + \min(x_2, \xi_{2,i})s_2] : x \in \mathbb{X}\}$



Robust optimization

Ben-Tal, Nemirovski,...
An additional constraint: overproduction should be avoided. We add the constraint

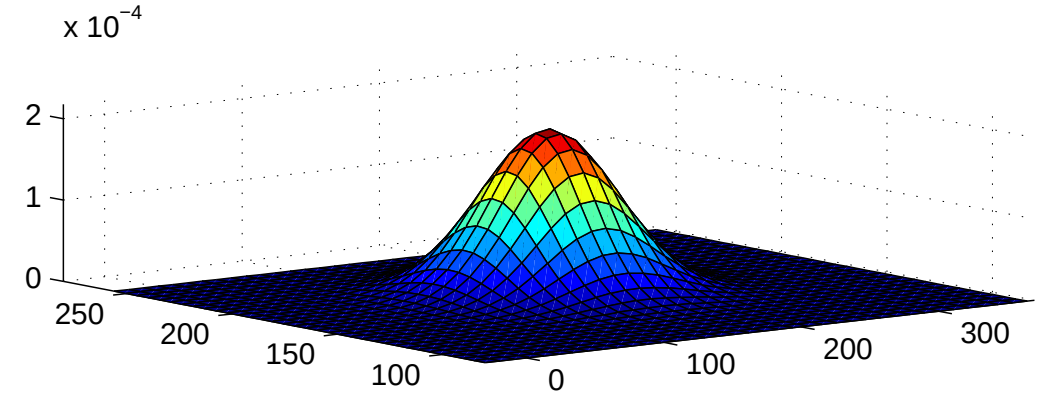
$$P\{x_1 > \xi_1 \text{ or } x_2 > \xi_2\} \leq 10\%$$

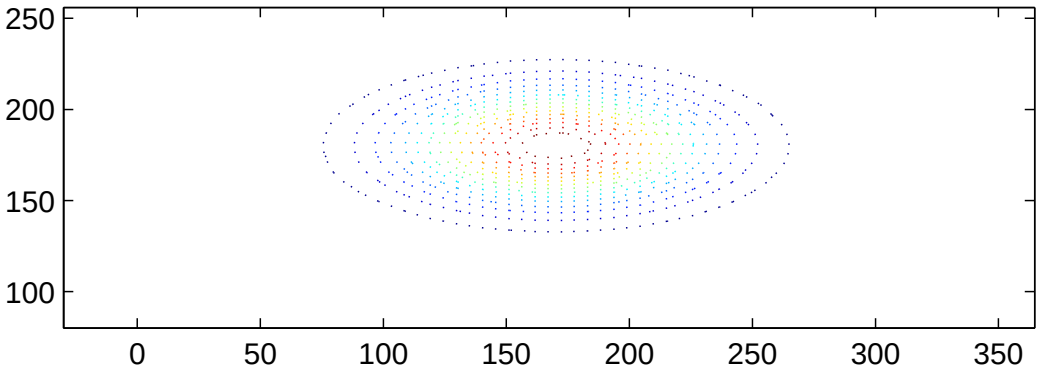
which is equivalent to

$$P\{x_1 \leq \xi_1 \text{ and } x_2 \leq \xi_2\} \geq 90\%$$

The scenario-robust method finds first an acceptance set A such that $P\{(\xi_1, \xi_2) \in A\} = 90\%$. and adds the constraints

$$x_1 \leq z_1; x_2 \leq z_2 \text{ for all } (z_1, z_2) \in A$$

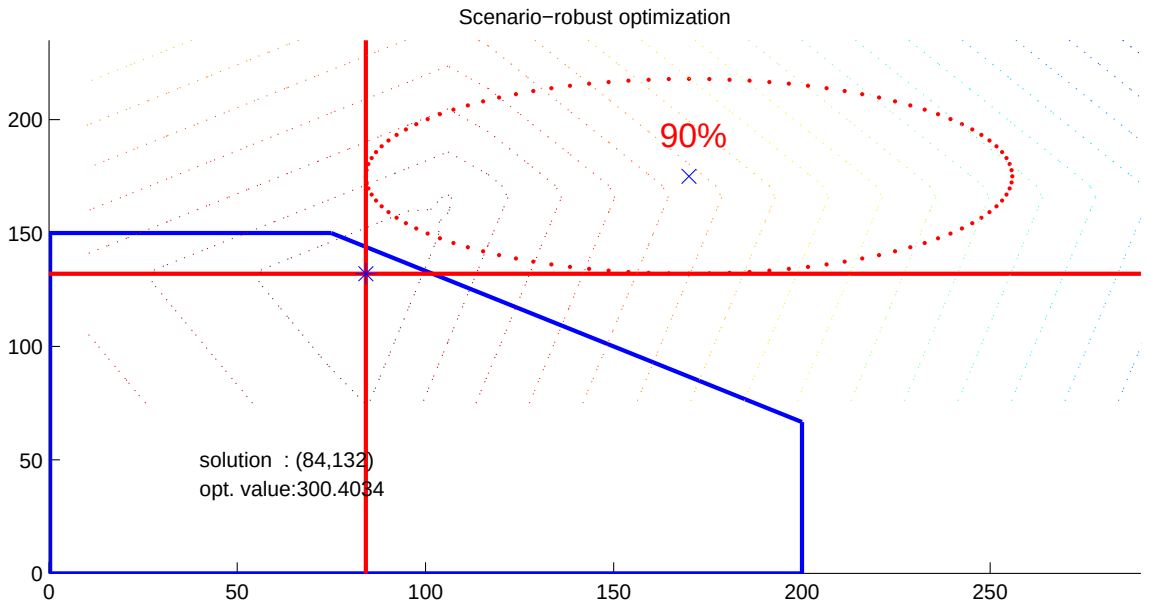

x 10^-4
2
1
0
250
200
150
100
0
100
200
300


250
200
150
100
0
50
100
150
200
250
300
350

Scenario-robust optimization is very pessimistic

Objective:

$$\max\{[\min(x_1, \xi_1)s_1 + \min(x_2, \xi_1)s_2] : x \leq z \text{ for } z \in A; x \in \mathbb{X}\}$$



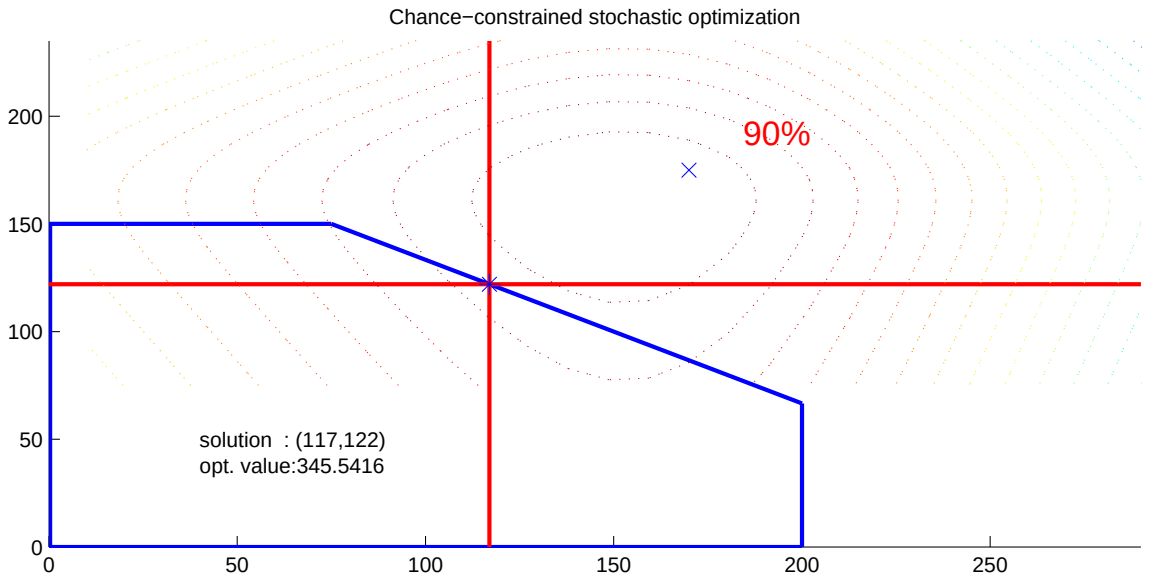
The wrong solution

Chance-constraint optimization is the correct way

The constraint

$$P\{x_1 \leq \xi_1 \text{ and } x_2 \leq \xi_2\} \geq 90\%$$

is added to the constraint set, but the acceptance set A may depend on the solution x and *is not chosen beforehand*.



The correct solution

Robust optimization-an additional example

Suppose that an investment can be made in 3 different assets and that 5 equally probable scenarios for the returns are given. The 5×3 data matrix

$$\begin{array}{ccc} \xi_1 & \xi_2 & \xi_3 \\ \end{array}$$
$$\left(\begin{array}{ccc} 1.10 & 0.96 & 0.96 \\ 1.08 & 1.06 & 1.05 \\ 1.02 & 1.06 & 1.05 \\ 0.98 & 1.01 & 1.00 \\ 1.00 & 1.00 & 0.90 \end{array} \right) \begin{array}{c} \omega_1 \\ \omega_2 \\ \omega_3 \\ \omega_4 \\ \omega_5 \end{array}$$

collects the 5 possible outcomes of the return variable $\xi = (\xi_{1,}, \xi_2, \xi_3)$.

The classical mean-variance (Markovitz-type) optimization problem reads

$$\begin{aligned} \min \quad & x^\top \mathbb{C}\mathrm{ov}(\xi)\, x \\ \text{subject to} \quad & \mathbb{E}(\xi)\, x \geq r_{min}, \\ & x \geq 0. \end{aligned} \tag{1}$$

Here, $\mathbb{E}(\xi)$ is the mean return vector (columnwise sums divided by 5) and $\mathbb{C}\mathrm{ov}(\xi)$ is the covariance matrix of ξ. The threshold r_{min} is a minimally required expected return, which we set to 1 here. Suppose in addition that we want to safeguard ourselves against large drops in portfolio value by requiring that the probability that the return is larger than 0.9750 (say) is at least 80 %. We would then add the constraint

$$P\left(\xi^\top x \geq 0.975\right) \geq 0.8. \tag{2}$$

This additional constraint reduces the feasible set.

In robust optimization, one would first choose a subset $\Omega' \subset \Omega$ of scenarios with probability $P(\Omega') = 0.8$ and require then the validity of

$$\xi(\omega)^\top x \geq 0.975 \qquad \text{for all } \omega \in \Omega'. \tag{3}$$

In our example, obviously 4 out of 5 scenarios should fulfill this inequality. Which scenario to be left out? Of course a bad one. It is reasonable to exclude the last row, i.e., to choose $\Omega' = \{\omega_1, \ldots, \omega_4\}$, since ξ_3 may drop to 0.9 in scenario ω_5, the overall worst value.

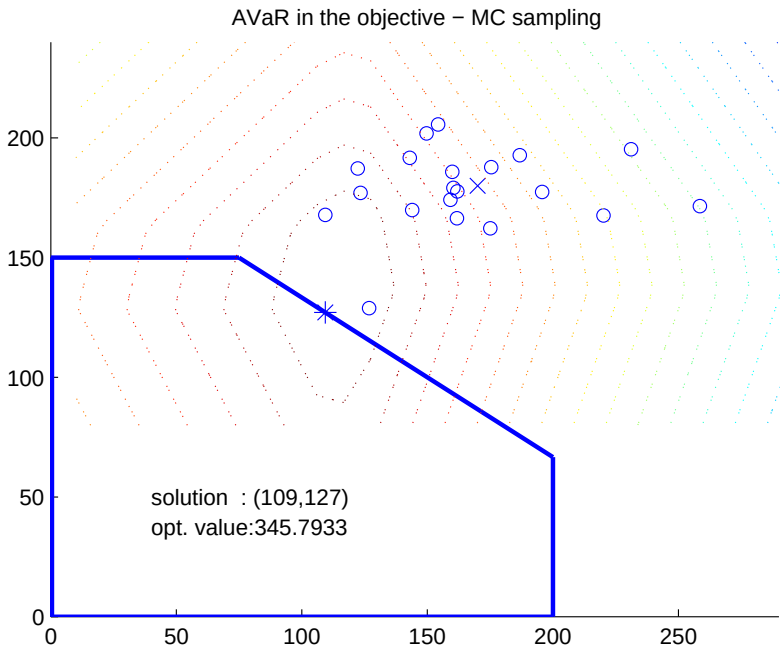
Solving the problem (1) with the additional constraint (3) leads to the solution $x^+ = (0.4031, 0.5731, 0)$. If, however, the full chance constrained stochastic problem (1) + (2) is solved, one gets a different and better solution $x^* = (0.4102, 0.5647, 0)$. Why can this happen? Simply because when choosing the bad scenario beforehand in the robust setup, one could not know that the optimal solution will not pick asset 3 at all ($x_3 = 0$) so that bad cases for asset 3 are irrelevant. If, however, one allows to choose the bad scenarios dependent on the decision, one finds that for the optimal decision x^* the scenario ω_4 is the worst and this one should form the exception set in the chance constrained stochastic problem.

Including risk aversion

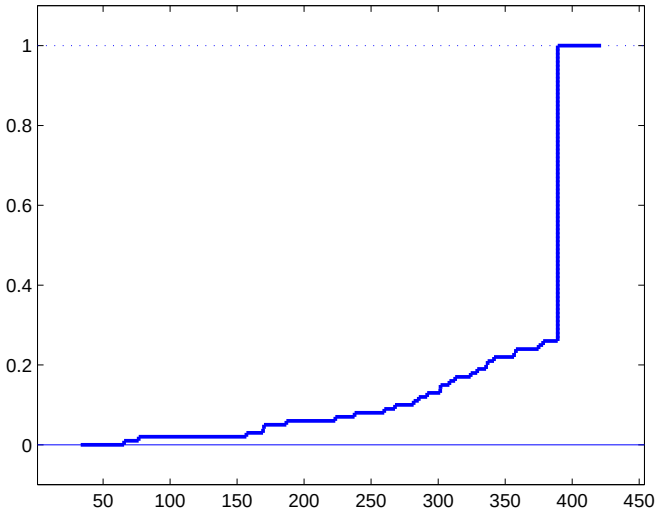
$$\mathbb{AV@R}_\alpha(Y) = \frac{1}{\alpha}\int_0^\alpha G_Y^{-1}(u)\,du \le \int_0^1 G_Y^{-1}(u)\,du = \mathbb{E}(Y)$$

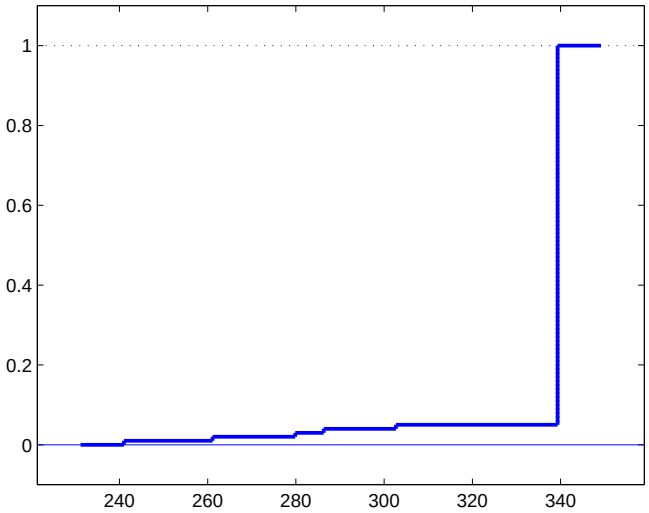
$\mathbb{AV@R}$ is an *utility functional* for a profit variable Y

Objective: $\max\{\mathbb{AV@R}_\alpha[\min(x_1,\xi_1)s_1 + \min(x_2,\xi_2)s_2] : x \in \mathbb{X}\}$



The influence of the risk objective to the profit distribution





Left: Objective = Expectation (risk neutral), Right: Objective = $\mathbb{A}$V@R (risk averse)

In the right picture, the downside risk is much smaller, but also the expected return is smaller.

Properties of risk functionals

We formulate here the risk for a cost/loss variable L: In the most general form, a *risk functional*

$$\rho_P(L|\mathcal{F})$$

has three arguments: The random cost/loss variable L, the probability measure P and the information, i.e. the sigma-algebra $\mathcal{F}$. Typical properties are:

- Convexity of $L \mapsto \rho_P(L)$
- Monotonicity of $L \mapsto \rho_P(L)$
- Positive homogeneity of $Y \mapsto \rho_P(L)$
- Concavity of $P \mapsto \rho_P(L)$
- Antimonotonicity of $\mathcal{F} \mapsto \rho_P(L|\mathcal{F})$

Utility and risk functionals/ profit and cost/loss variables

If Y is a profit variable, then $L = -Y$ is a loss variable and vice versa.
If $\mathcal{U}$ is an utility functional, then $\rho(\cdot) = -\mathcal{U}(\cdot)$ is a risk functional.
For profit variables, we maximize $\mathcal{U}(Y)$ and minimize $\rho(Y)$.
For cost/loss variables, we maximize $\mathcal{U}(-L) = \mathcal{U}(Y)$ and minimize $\rho(-L) = \rho(Y)$. Notice that for cost/loss variables, all properties remain the same, except that monotonicity and antimonotonicity are reversed.

Ambiguity

According to Ellsberg (1961) we face 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" - Knightian uncertainty according to F. Knight "Risk, Uncertainty and Profit" (1920)).

Coping with model error

In many applications, the model is identified based on data and therefore the problem is subject to *model error*. Instead of the estimated baseline model $\hat{P}$, some other probability models P may also be compatible with the data and describe the real phenomenon well.
Therefore we define an ambiguity set of models $\mathcal{P}$ and solve a *minimax ambiguity model*

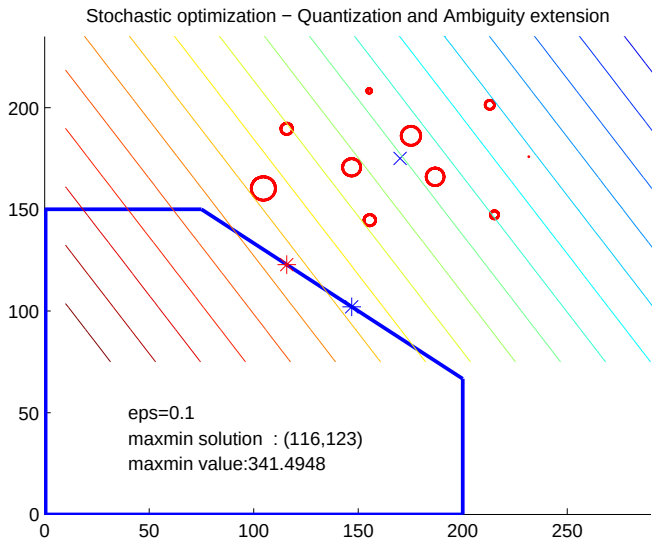
$$\max \min_{P\in\mathcal{P}} \{\mathbb{E}_P[\min(x_1,\xi_1)s_1 + \min(x_2,\xi_2)s_2 - c_1x_1 - c_2x_2] : x \in \mathbb{X}\}.$$

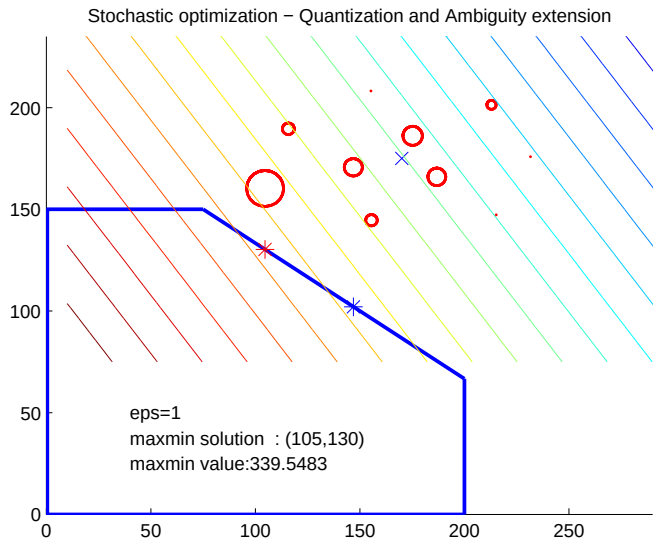
The ambiguity problem is a maximin problem and solved by algorithms for finding a saddlepoint (x^*, P^*). P^* is the *worst case model*.
We choose typically the ambiguity set as a ball around the baseline model P:

$$\mathcal{P} = \{P : d(\hat{P}, Q) \leq \epsilon\}$$

where ϵ is the *ambiguity radius*.

Ambiguity





Left: Ambiguity radius $\epsilon = 0.1$ Right: Ambiguity radius $\epsilon = 1$

Cost of ambiguity and reward for distributional robustness

Let

$$F(x,P) = \mathbb{E}_P[\min(x_1,\xi_1)s_1 + \min(x_2,\xi_2)s_2 - c_1x_1 - c_2x_2]$$

be the our objective problem. Let (x^*, P^*) be the saddle point. Then

$$F(x,P^*) \le F(x^*,P^*) \le F(x^*,P).$$

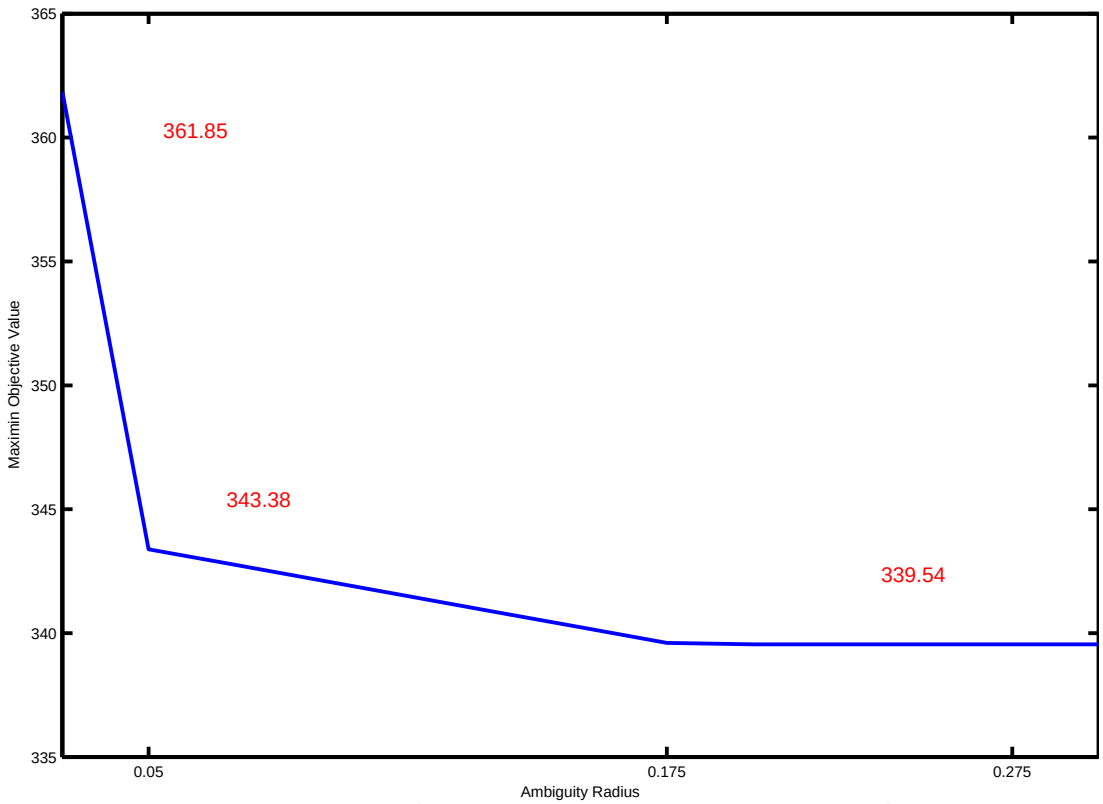
Let $\hat{P}$ be our baseline model and $\hat{x}$ the pertaining optimal solution. The cost of ambiguity (COA):

$$F(\hat{x},\hat{P}) - F(x^*,\hat{P}).$$

The reward for distributional robustness (RDR):

$$F(x^*,P^*) - F(\hat{x},P^*).$$

The cost of ambiguity

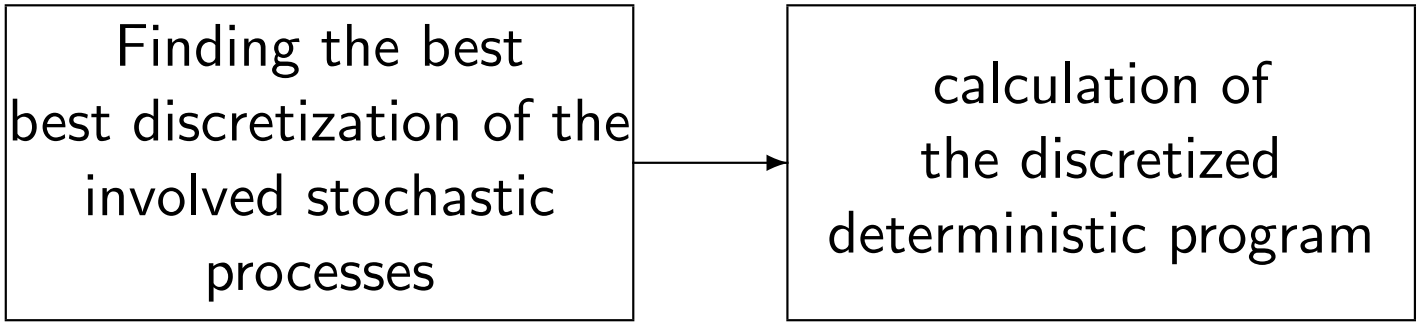


The drop in optimal value (the maximin value) as a function of the ambiguity radius ϵ.

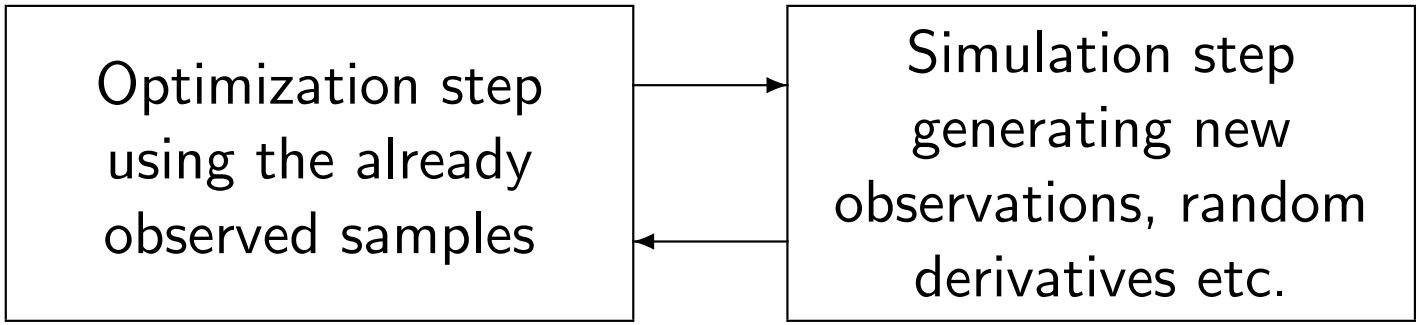
Summary

- Stochastic optimization allows to deal with uncertainties in decision parameters.
- The VSS and the EVPI values characterize the degree of "stochasticity" of the problem.
- To numerically solve a stochastic program, an approximation scheme, such as MC, QMC or quantization is typically required.
- We distinguish between methods, where the approximation is done beforehand and kept fixed, or methods where the approximation is interleaved with the optimization. Some deterministic optimization problems are solved by introducing artificial random variables (random search, ACO, genetic algorithms). We do not consider these as stochastic optimization problems.

Approximation by discretization



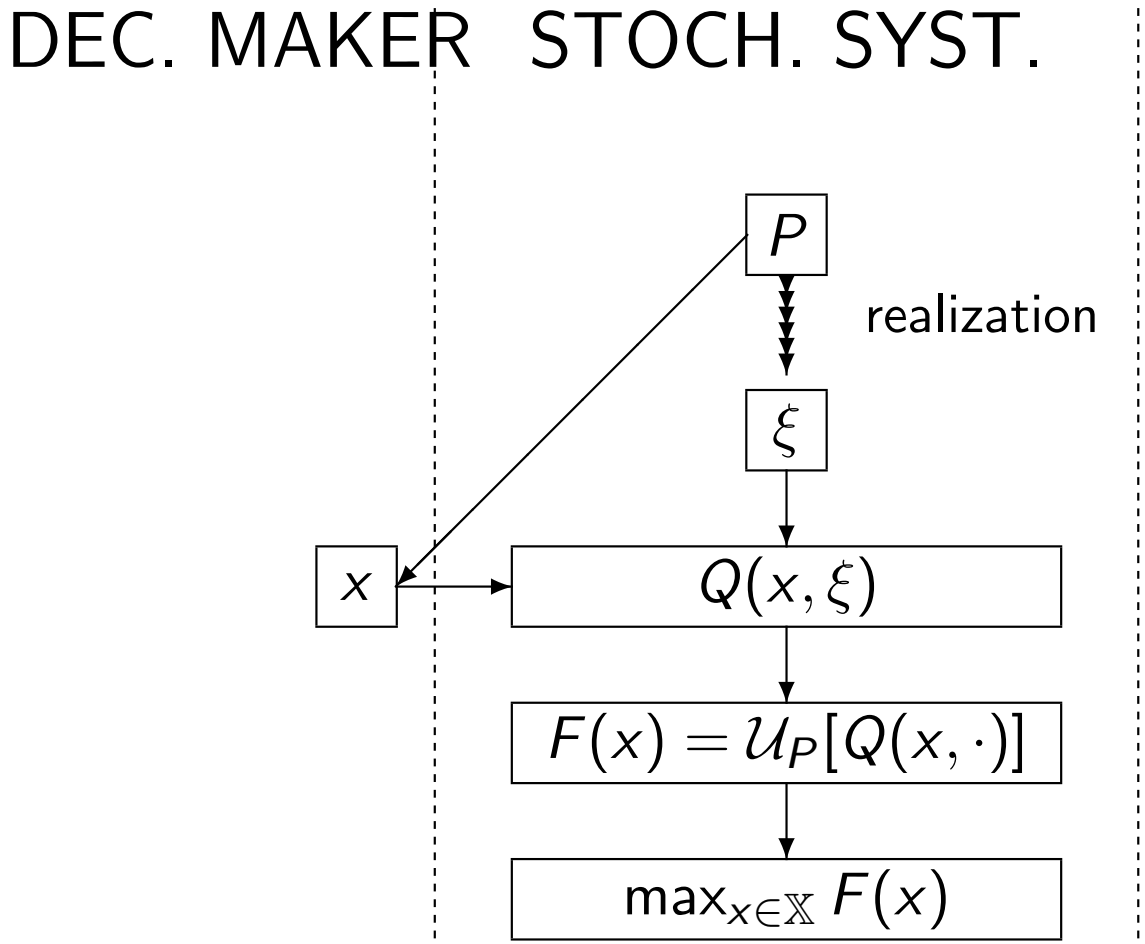
Optimization interleaved with Simulation



Examples: Stochastic Approximation (SA), Sample Average Approximation (SAA), Stochastic Dual Dynamic Programming (SDDP)

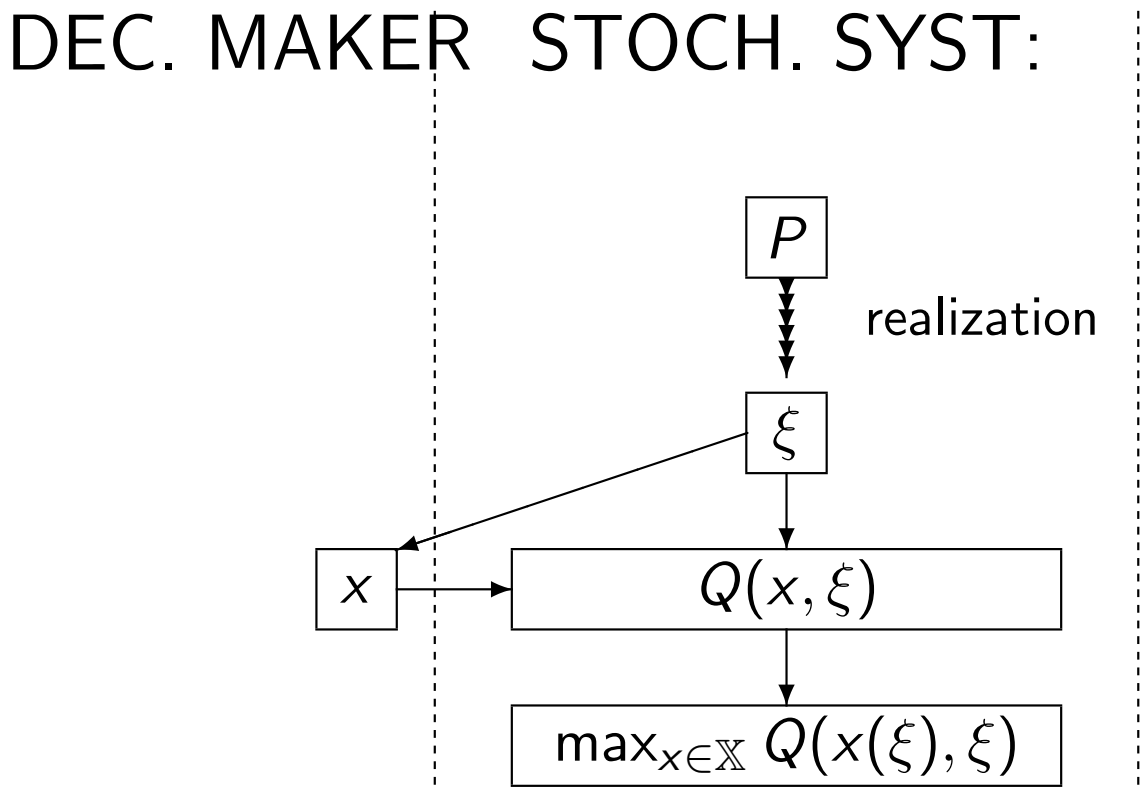
- Robust optimization chooses first a typical set of outcomes and looks then for the worst case, while for problems with probability constraints, the exception set is determined as part of the optimization.
- Risk functionals allow to formulate risk aversion.
- More sophisticated decision problems are stochastic multi-stage. For these problems, the amount of information available at each stage is quite relevant.
- Model uncertainty (ambiguity) requires distributionally robust decisions, which can be found by solving maximin (minimax) problems. We look at the COA *Cost of ambiguity* versus the RDR *reward for distributionally robustness*

Stochastic optimization: here and now



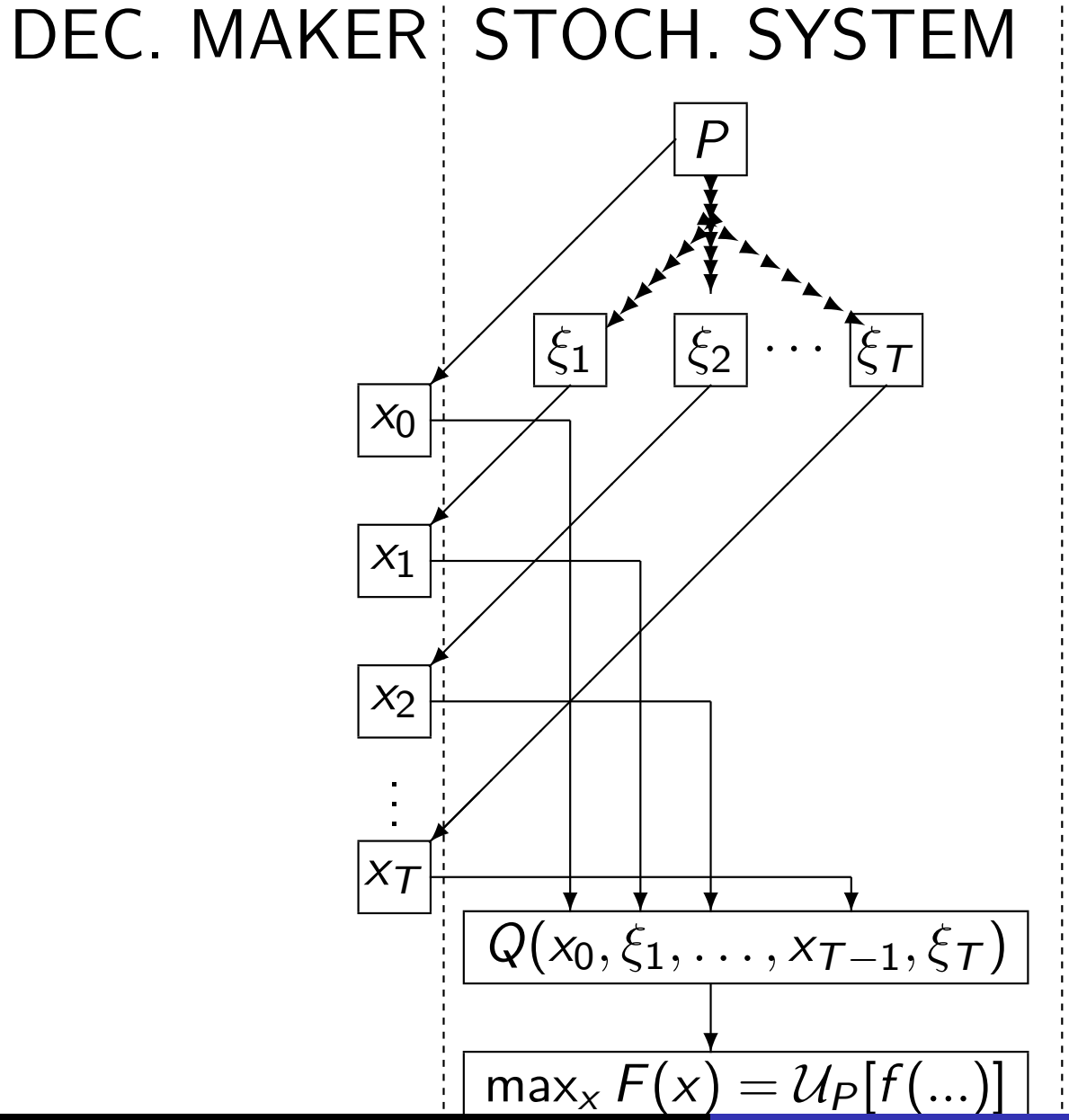
$\mathcal{U}_P$ is an utility functional, like expectation, risk corrected expectation or a distortion functional, e.g. the average value-at-risk.

Stochastic optimization: wait and see

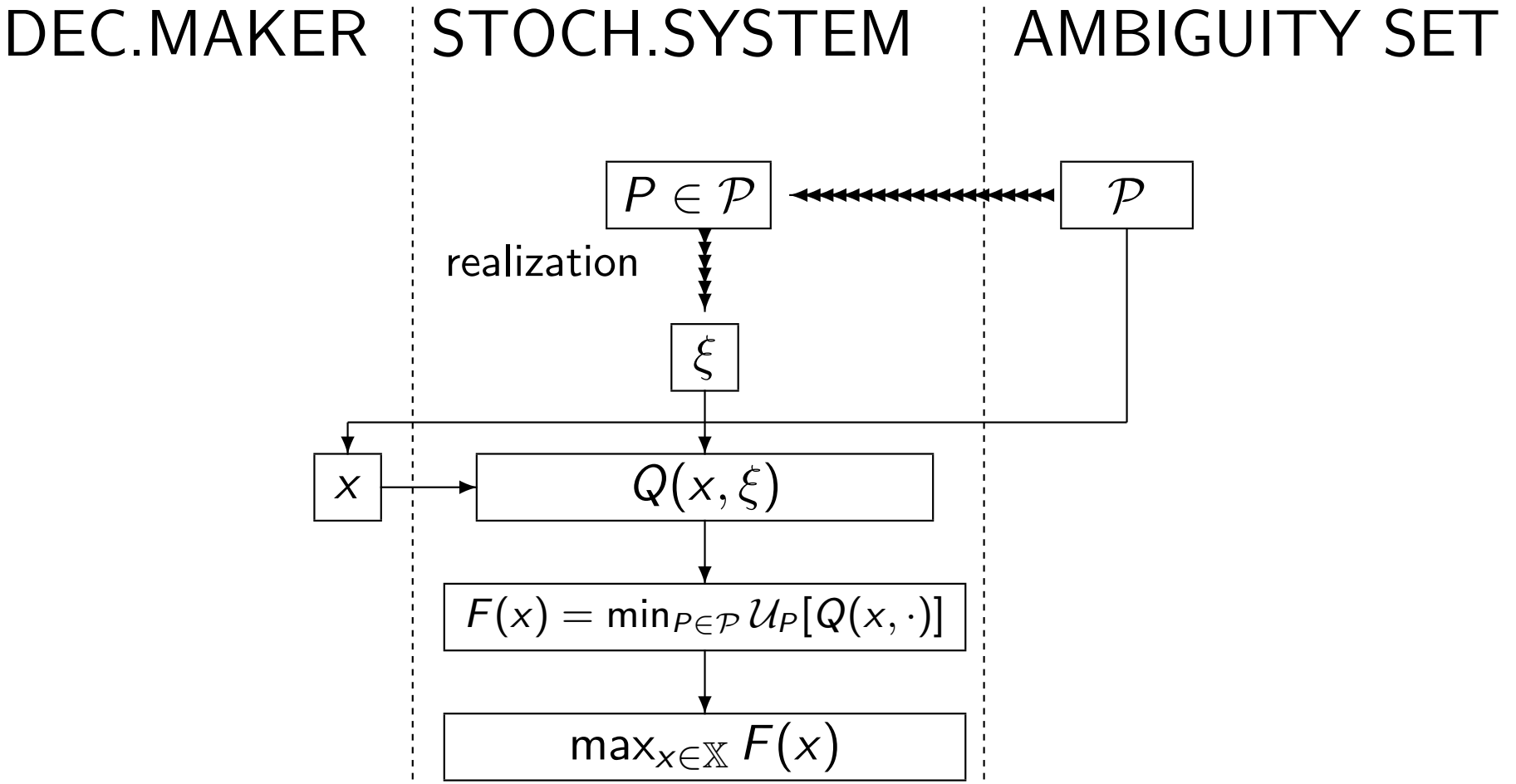


This problem decomposes scenariowise.

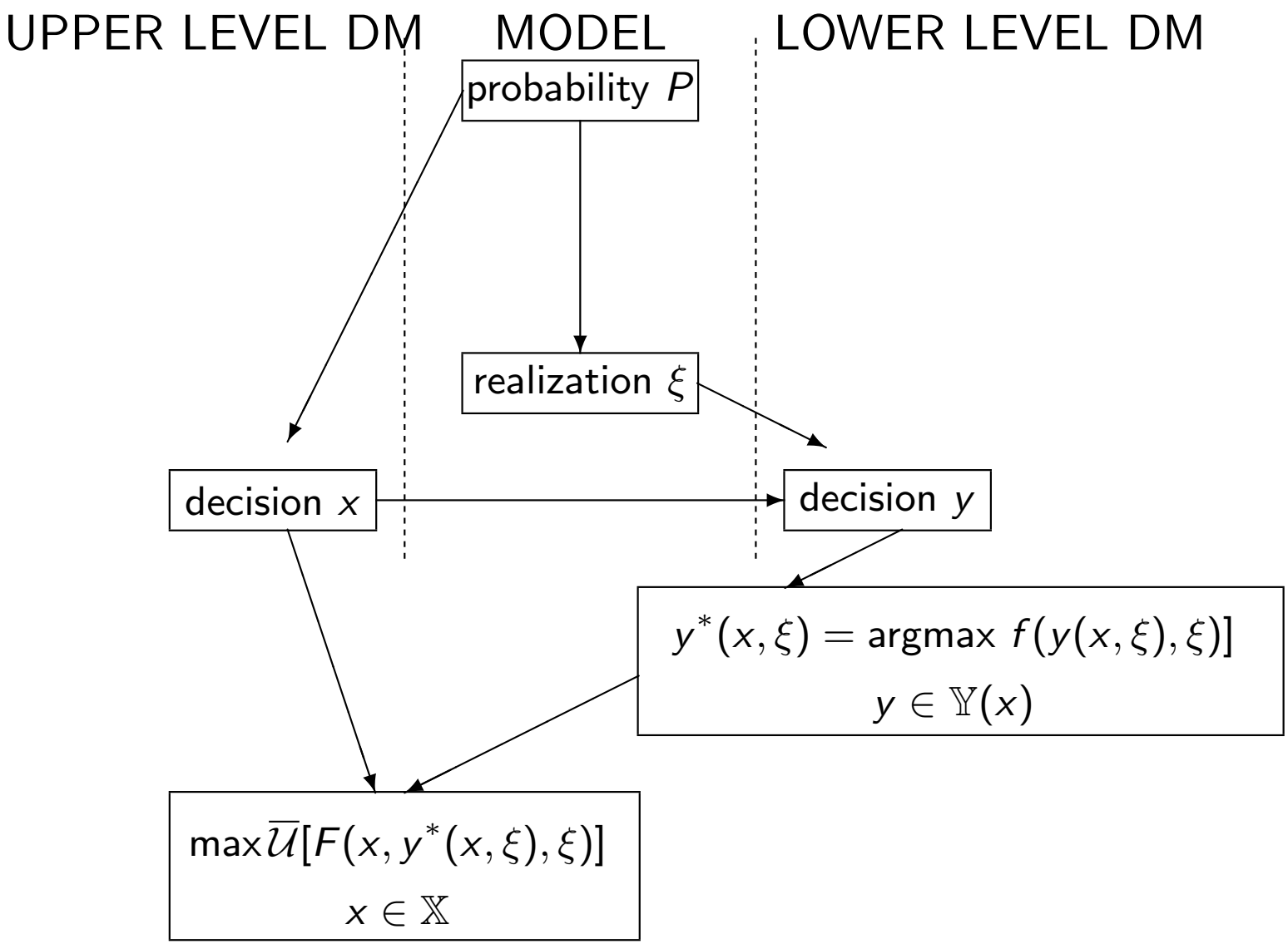
Multistage stochastic optimization



Stochastic optimization under ambiguity



Bilevel problems; UL: here and now, LL: (nearly) wait and see



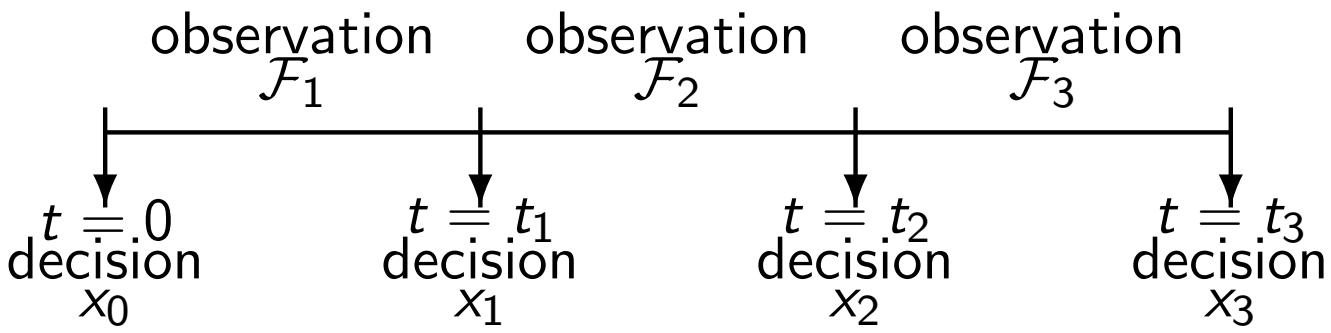
Formulation of a multistage stochastic program

Problem $(\mathcal{P})$: $\min\{\rho[Q(x_0, \xi_1, \ldots, x_{T-1}, \xi_T)] : x \lhd \mathfrak{F}\}$,

where

$\xi = (\xi_1, \ldots, \xi_T)$	a random scenario process, on a probability space $(\Omega, \mathcal{F}, P)$
$x = (x_0, \ldots, x_{T-1})$	the sequence of decisions,
$Q(x_0, \xi_1, \ldots, \xi_T)$	the profit function,
ρ	a version-independent risk functional, such as the expectation $\mathbb{E}$
$\mathfrak{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_T)$	a filtration (an increasing sequence of σ-algebras) on $(\Omega, \mathcal{F}, P)$
$\xi \lhd \mathfrak{F}$	ξ is adapted to $\mathfrak{F}$, i.e. $\sigma(\xi) \subseteq \mathfrak{F}$
$x \lhd \mathfrak{F}$	the nonanticipativity condition.

Non-anticipativity



Types of stochastic programs

- Single-stage stochastic programs

  decision $x_0 \to$ observation ξ_1

- Two-stage stochastic programs

  decision $x_0 \to$ observation $\xi_1 \to$ recourse decision x_1

  decision $x_0 \to$ observation $\xi_1 \to$ decision $x_1 \to$ observation ξ_2

- Multistage stochastic programs

  decision $x_0 \to$ observation $\xi_1 \to \ldots \to$ decision x_T

Types of stochastic programs

- If the functional is the expectation, we call it a *risk-neutral* problem
- If the functional models the risk or the acceptability (utility), it is called a *risk-sensitive* problem
- Some stochastic programs are formulated with the help of *multiperiod risk functionals* for intermediate risk:

$$\min\{\mathcal{R}[Q_1(x_0,\xi_1); Q_2(x_1,\xi_2); \dots, Q(x_{T-1},\xi_T)]$$

- If the probability of an event is constrained, we call it a *chance-constrained* problem
- Some stochastic programs (but not all) allow formulation as a *dynamic* program
- If the stochastic program is linear, we distinguish between models with *random right hand side*, *random costs* and/or *random technology matrix*.

Example: A multistage stochastic program

$$\min \{c_0(x_0) + \mathbb{E}_{\xi_1}[\min\ c_1(x_1,\xi_1) + \mathbb{E}_{\xi_2}[\min\ c_2(x_2,\xi_2) + \\ + \mathbb{E}_{\xi_T}[\cdots + \mathbb{E}[\min\ c_T(x_T,\xi_T)]\ldots]]] : x \in \mathbb{X}\},$$

where the feasible set $\mathbb{X}$ is given by

$$\begin{aligned} W_0 x_0 &\geq h_0 \\ A_1 x_0 + W_1 x_1 &\geq h_1(\xi_1) \\ A_2 x_1 + W_2 x_2 &\geq h_2(\xi_2) \\ &\vdots \\ A_T x_{T-1} + W_T x_T &\geq h_T(\xi_T) \qquad (4) \\ x_1 &\lhd \mathcal{F}_1 \\ &\vdots \\ x_T &\lhd \mathcal{F}_T. \end{aligned}$$

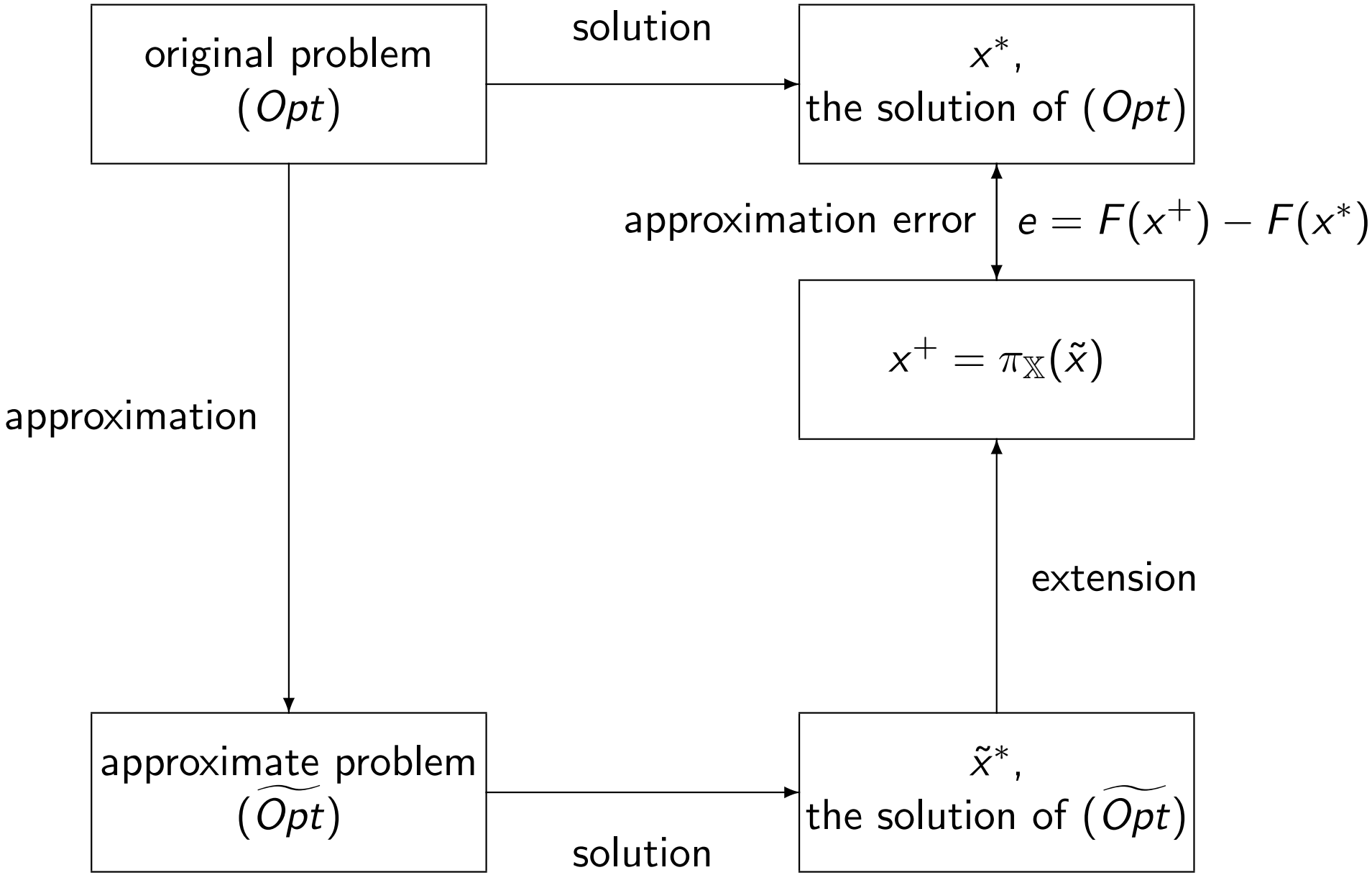
A: recourse matrix; W: technology matrix, h: right hand side

Solution through finite approximation

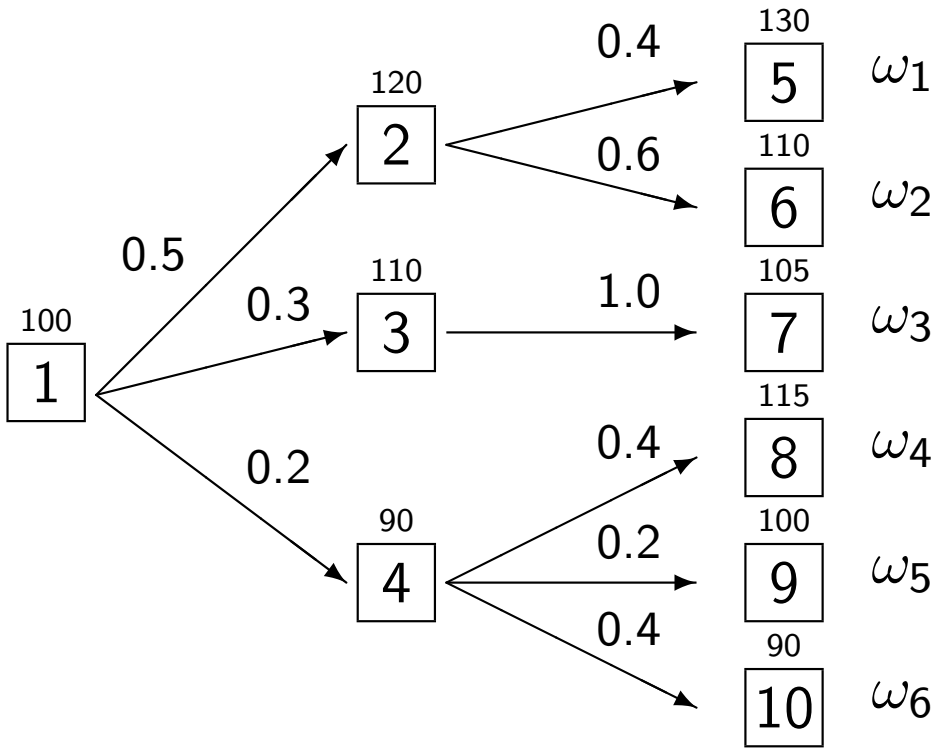
Instead of the process $\xi = (\xi_1, \dots, \xi_T)$ and the filtration $\mathfrak{F} = (\mathcal{F}_1, \dots, \mathcal{F}_T)$ we consider a simpler process $\tilde{\xi} = (\tilde{\xi}_1, \dots, \tilde{\xi}_T)$ and the filtration $\tilde{\mathfrak{F}} = (\tilde{\mathcal{F}}_1, \dots, \tilde{\mathcal{F}}_T)$.
Filtrations on a finite probability spaces are equivalent to trees (of subsets).
The basic problem is replaced by the simpler problem

$$\text{Problem } (\tilde{\mathcal{P}}): \ \min\{\mathcal{R}[Q(x_0, \tilde{\xi}_1, \dots, x_{T-1}, \tilde{\xi}_T)] : x \lhd \tilde{\mathfrak{F}}\},$$

$->$ tree structured problem!


original problem
(Opt)
solution
x^*,
the solution of (Opt)
approximation error
$e = F(x^+) - F(x^*)$
$x^+ = \pi_{\mathbb{X}}(\tilde{x})$
approximation
extension
approximate problem
$(\widetilde{Opt})$
solution
$\tilde{x}^*$,
the solution of $(\widetilde{Opt})$

Valuated trees as basic data objects



ω_1 $P(\omega_1) = 0.2$

ω_2 $P(\omega_2) = 0.3$

ω_3 $P(\omega_3) = 0.3$

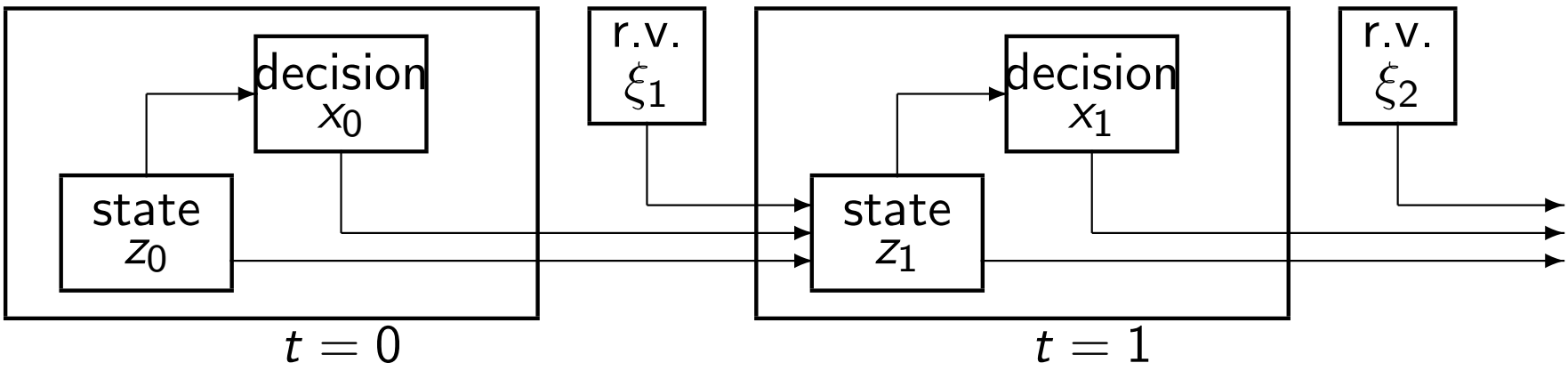
ω_4 $P(\omega_4) = 0.08$

ω_5 $P(\omega_5) = 0.04$

ω_6 $P(\omega_6) = 0.08$

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

Dynamic optimization



The decision dynamics

We may represent the multistage dynamic decision model as a state-space model. We assume that there is a state vector ζ_t, which describes the situation of the decision maker at time t immediately before he must make the decision x_t, for each $t = 1, \ldots, T$, and its realization is denoted by z_t. The initial state $\zeta_0 = z_0$, which precedes the deterministic decision at time 0, is known and given by ξ_0. To assume the existence of such a state vector is no restriction at all, since we may always take the whole observed past and the decisions already made as the state:

$$\zeta_t = (x^{t-1}, \xi^t), \qquad t = 1, \ldots, T.$$

Here $x^t = (x_1, \ldots, x_t)$ and $\xi^t = (\xi_1, \ldots, \xi_t)$. However, the vector of required necessary information for the future decisions is often much shorter.

The state variable process $\zeta = (\zeta_1, \dots, \zeta_T)$, with realizations $z = (z_1, \dots, z_T)$, is a controlled stochastic process, which takes values in a state space $Z = Z_1 \times \cdots \times Z_T$. The control variables are the decisions x_t, $t = 1, \dots, T$. The state ζ_t at time t depends on the previous state ζ_{t-1}, the decision x_{t-1} following it, and the last observed scenario history ξ^t. A transition function g_t describes the state dynamics:

$$\zeta_t = g_t(\zeta_{t-1}, x_{t-1}, \xi^t), \qquad t = 1, \dots, T. \tag{5}$$

At the terminal stage T, no decisions are made, only the outcome $\zeta_T = z_T$ is observed.

Note that ζ_t is a random variable with realization z_t, which is a function of the realization of ξ^t, for each $t = 1, \ldots, T$.
The decision x of the multistage stochastic problem is a vector of functions $x = (x_0, \ldots, x_{T-1})$, $t = 1, \ldots, T-1$, maps Ξ^t to $\mathbb{R}^{m_t}$.
We require that the feasible decision x_t at time t satisfies a constraint of the form

$$x_t \in \mathcal{X}_t(\zeta_t), \qquad t = 1, \ldots, T,$$

where $\mathcal{X}_t$ are closed convex multifunctions with closed convex values.

Bellmann equation

If the problem decomposes into a sum of profit functions and the probability functional is the expectation, one may use a stage-wise decomposition.
For each stage, one defines a value function $V_t(z_t)$. Notice that z_t has to carry all relevant information for the future.
Backward equation:

$$V_t(z_t) = \min_{x_t \in \mathcal{X}_t} \mathbb{E}[Q_t(x_t, \zeta_t, \xi_{t+1}) + V_{t+1}(\zeta_{t+1})]$$

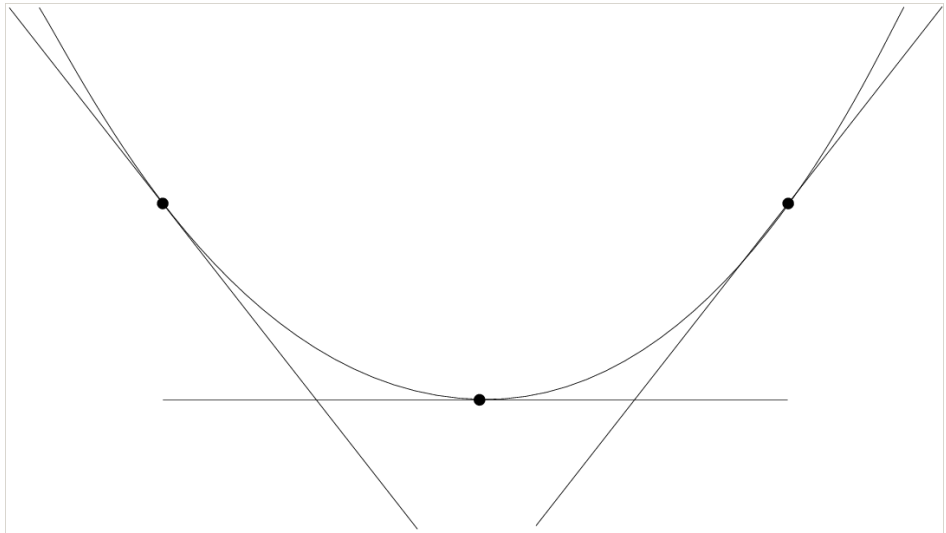
with

$$\zeta_{t+1} = g_{t+1}(\zeta_t, x_t, \xi^{t+1}).$$

Typically the functions $V_t(z)$ are discretized by using linear, quadratic or cubic interpolations between finitely many support points.

Subgradient approximation

Convex functions can be approximated from below by every finite collection of subgradients.



Benders decomposition

$$\left\| \begin{array}{l} \text{Minimize } f(x) + g(y) \\ \text{s.t.} \\ Tx + Ay \geq b \\ x \in S \\ y \geq 0 \end{array} \right.$$

Here x and y are vectors and T and A are matrices of appropriate sizes. We call x the first stage and y the second stage variables. Let $\mathcal{X} = \{x : \exists y \geq 0 \text{ such that } Tx + Ay \geq b\}$ be the implied feasibility set. $\mathcal{X}$ is a convex polyhedron. We will find representation for $\mathcal{X}$.

Let $\mathcal{H} = \{h : \exists y \geq 0 \text{ such that } Ay \geq h\}$. $\mathcal{H}$ is also a convex polyhedron. $h \in \mathcal{H}$ implies that the following linear program is feasible

$$\left\| \begin{array}{l} \text{Minimize } 0^\top y \\ \text{s.t.} \\ Ay \geq h \\ y \geq 0 \end{array} \right.$$

and its optimal value is 0, which is equivalent to the fact that its dual is bounded

$$\left\| \begin{array}{l} \text{Maximize } h^\top u \\ \text{s.t.} \\ A^\top u \leq 0 \\ u \geq 0 \end{array} \right.$$

and its optimal value is also 0.

This shows that $\mathcal{H}$ can be represented in the following way: Let $\mathcal{U} = \{u | A^\top u \leq 0, u \geq 0\}$. Then $\mathcal{H} = \{h : h^\top u \leq 0 : u \in \mathcal{U}\}$. Let $(u_i)_{i \in I}$ be the set of extremals of $\mathcal{U}$. Then
$\mathcal{H} = \{h : h^\top u_i \leq 0 : i \in I\}$ and finally
$\mathcal{X} = \{x : b - Tx \in \mathcal{H}\} = \{x : x^\top T^\top u_i \geq b^\top u_i : i \in I\}$. Let $w_i = T^\top u_i$ and $\nu_i = b^\top u_i$ and let W be the matrix with rows $w_i^\top$ and v be the vector with components v_i. Then

$$\mathcal{X} = \{x : Wx \geq v\}.$$

Let further

$$Q(x) = \min\{g(y) : Ay \geq b - Tx; y \geq 0\}.$$

For $x \notin \mathcal{X}$, we set $Q(x) = \infty$.
If g is convex, then $Q_1(h) = \min\{g(y) : Ay \geq h; y \geq 0\}$ is convex on $\mathcal{H}$ and therefore $Q(x) = \min\{g(y) : Ay \geq b - Tx; y \geq 0\}$ is convex on $\mathcal{X}$. As every convex function, Q coincides with the maximum of (possibly infinitely many) linear functions on R.

If g is linear, say $g(y) = d^\top y + \gamma$, then Q is a convex, piecewise linear function. To see this, let $Q_1(h) = \min\{d^\top y + \gamma : Ay \geq h; y \geq 0\}$. By duality, $Q_1(h) = \gamma + \max\{z^\top h : A^\top z \leq d; z \geq 0\}$. The polyhedron $\{z : A^\top z \leq d; z \geq 0\}$ is the convex hull of its extremals $(z_k)_{k\in K}$. Therefore $Q(x) = \gamma + \max\{-z_k^\top Tx + z_k^\top b : k \in K\}$. Let $r_k = -T^\top z_i$ and $\tau_k = z_k^\top b$. Then

$$Q(x) = \gamma + \max\{r_k^\top x + \tau_k : k \in K\}.$$

If g is the maximum of linear functions, say
$g(y) = \max\{d_\ell^\top y + \gamma_\ell : \ell \in L\}$, then
$Q_1(h) = \min\{\xi : \xi \geq d_\ell^\top y + \gamma_\ell\} : Ay \geq h; y \geq 0\}$. Let D be the matrix consisting of the rows d_ℓ and g the vector with components γ_ℓ. By duality,
$Q_1(h) = \max\{z^\top h + v^\top g : A^\top z \leq Dv; v^\top \mathbb{1} = 1; v \geq 0; z \geq 0\}$.
The polyhedron $\{(v, z) : A'z \leq Dv; v^\top \mathbb{1} = 1; v \geq 0; z \geq 0\}$ is the convex hull of its extremals $(v_k, z_k)_{k \in K}$. Let $r_k = -T^\top z_i$ and $\tau_k = z_k^\top b + v_k^\top g$. Then

$$Q(x) = \max\{r_k^\top x + \tau_k : k \in K\}.$$

Therefore $Q(x) = \max\{-z_k^\top Tx + z_k^\top b : k \in K\}$.

We call $\mathcal{X}^{(s)}$ an outer approximant of $\mathcal{X}$, if it consists only of a subset of constraints of $\mathcal{X}$. We call $Q^{(s)}$ a lower approximant of Q, if it is the maximum of a subset of the functions appearing in Q.
Here is the structure of the algorithm.
Let $\mathcal{X}^{(1)}$, $Q^{(1)}$ be some simple approximants of $\mathcal{X}$ and Q.

1. Set $s := 1$.
2. [Master] Solve $\min\{f(x) + Q^{(s)}(x) : x \in S, x \in \mathcal{X}^{(s)}\}$. Send x and $Q^{(s)}$ to slave.
3. [Slave] Solve $\min\{g(y) : Ay \geq b - Tx; y \geq 0\}$.
   If this program is feasible then $x \in \mathcal{X}$. If moreover $Q(x) = Q^s(x)$, we have found a solution and stop.
   - 3.1 If the Slave program is not feasible, then we have to find a constraint $w_i x \geq \nu_i$, which is not satisfied. Send this constraint (w_i, ν_i) ("feasibility cut") to the master.
   - 3.2 If the Slave program is feasible, but $Q^{(s)}(x) < Q(x)$, then we calculate a supporting hyperplane to Q in the point x: Let $r \in \partial Q(x)$ be a subgradient of Q at x and let $\tau = Q(x) - r'x$. Then for all z, $Q(z) \geq r'z + \tau$, with equality for $z = x$. Send r and τ ("optimality cut") to the master.
4. [Master] The master adds the feasibility cut to R^s to get $R^{(s+1)} = R^{(s)} \cap \{x : w_i x \geq \nu_i\}$ and/or updates $Q^{(s)}$ using the optimality cut $Q^{(s+1)} = \max[Q^{(s)}, r'x + \tau]$.
5. goto 2.

Since we add always new cuts and do not forget old ones and since R is a polyhedron with finitely many extremals and Q is convex piecewise linear, we find a solution in finitely many steps.
The calculation of feasibility cuts: Suppose that there is no $y \geq 0$ such that $Ay \geq b - Tx$. This means that the problem

$$\left\| \begin{array}{l} \text{Minimize } 0'y \\ \text{s.t.} \\ Ay \geq b - Tx \\ y \geq 0 \end{array} \right.$$

is infeasible and hence its dual

$$\left\| \begin{array}{l} \text{Maximize } (b - Tx)'u \\ \text{s.t.} \\ A'u \leq 0 \\ u \geq 0 \end{array} \right.$$

is unbounded. One may identify an extreme ray u, along which the problem is unbounded, i.e. $u \geq 0$, $A'u \leq 0$, $(b - Tx)'u \geq 0$, $(b - Tx)'u \neq 0$. Let $w = T'u$, $\nu = b'u$. The pair (w, ν) gives the new constraint $w'x \geq \nu$ for $\mathcal{X}$.

The calculation of optimality cuts: Solving $\min\{g(y) : Ay \geq b - Tx; y \geq 0\}$ leads to a pair (y, λ) of primal and dual solutions. The subgradient $r \in \partial Q(x)$ is $r = -T'\lambda$. The very same procedure applies to convex, nonlinear f and g with the only difference, that the procedure does not terminate in finitely many steps.

The same algorithm applies to the tree structured case, where the second stage decomposes completely into the sum of J functions:

$$\left\| \begin{array}{ll} \text{Minimize } f(x) + \sum_{j=1}^{J} g_j(y_j) & \\ \text{s.t.} & \\ T_j x + A_j y_j \geq b_j & \\ x \in S & \\ y_j \geq 0 & \text{for all } j \end{array} \right.$$

This problem is is solved by one master program and J slave programs: Each of the J slaves is responsible for one of the functions g_j. Notice that these functions are only linked through the first stage variables x. Therefore, the slave optimizations can be run in parallel: The master uses approximations $\mathcal{X}^{(s)}$ and $Q_j^{(s)}$ of the functions $Q_j(x) = \min\{g_j(y) : A_j y \geq b - T_j x; y \geq 0\}$. It solves $\min\{f(x) + \sum_{j=1}^{J} Q_j^{(s)}(x) : x \in S, x \in \cap_j \mathcal{X}_j^{(s)}\}$. It sends x and $Q_j^{(s)}$ to slave s and receives from him feasibility cuts for $\mathcal{X}_j$ and Q_j.

In the tree structure, every node which is not the root and not a leaf is at the same time master and slave. The root is only master, the leaf nodes are only slaves. Each nonterminal node n must optimize vector x_n of local decisions using its local constraints $x_n \in S_n$, the implied feasibility constraints $x_n \in \cap_{j \in n+} \mathcal{X}_j$. The function to be optimized is the sum of a local objective and the optimal value functions of the successor nodes $f_n(x_n) + \sum_{j \in n+} Q_j(x_n)$.

SDDP- Pereira

The SDDP approach is identical to the Benders decomposition, but does not construct the scenario tree right from the beginning, but uses values of the process which are sampled in during the procedure.
If the stochastic process ξ is independent, then this method is very efficient, since no conditional distributions appear. However, the independence assumption is very questionable and to modify the process in such a way that only independent innovations appear is not always possible.

The sample average approximation-SAA

This name is just another name for Monte Carlo Simulation in connection with stochastic optimization. Originally, it was just used for two-stage programs with expectation objective.
The program

$$v^* = \min_{x, y(\xi)} c(x) + \mathbb{E}_P[f(y(\xi), x, \xi)] : h(x, y, \xi) \geq 0$$

with x being the first stage decision and $y(\xi)$ are the second stage (recourse) decisions, is transformed into

$$\tilde{v} = \min_{x, y_i} c(x) + \frac{1}{n} \sum_{i=1}^{n} [f(y_i, x, \xi_i)] : h(x, y_i, \xi_i) \geq 0,$$

where $(\xi_1, \ldots, \xi_n)$ is an i.i.d. sample from P.
Notice that every solution (x^*, y_i^*) depends on the sample and is therefore a *random* solution. Also the minimal value is a random variable.

Research questions...

Much work has been devoted to SAA for two-stage problem to answer the questions

1. Do the minimal values $\tilde{v}$ of the SAA converge to the true v^* for $n \to \infty$?
2. Do the the random solutions converge to the true solutions for $n \to \infty$?
3. How fast is the convergence in 1.?
4. How fast is the convergence in 2.?
5. Is there asymptotic normality in 1. and can one get confidence intervals for the true solution?
6. Is there asymptotic normality in 2. and can one get confidence intervals for the true solution?

... and answers

1. yes.
2. yes, under uniqueness assumption.
3. $n^{-1/2}$ typically.
4. $n^{-1/2}$, but may be faster, if the set of constraints is not smooth at the solution.
5. In smooth cases yes, otherwise no.
6. can be complicated

The basic inequality for SAA

If we solve the problem for each scenario ξ_i separately, i.e.

$$v_i^+ = \min_{x,y} c(x) + [f(y,x,\xi_i)] : h(x,y,\xi_i) \geq 0$$

then

$$\frac{1}{n}\sum_{i=1}^{n} v_i^+ \leq \tilde{v}.$$

since for each sample, one may choose another first stage decision x, but there should be the same x for all samples.
With a similar argument

$$\mathbb{E}(\tilde{v}) \leq v^*.$$

Consequently, by repeating the solution procedure k times with always a new draw of the random sample of size n, one gets a valid lower bound.
A valid upper bound can always be found by choosing one feasible solution.

Extensions

The SAA method may also handle other functionals than the expectation, since we just replace the probability P by its sampled counterpart $\hat{P}_n$.
There is also a version for multistage, where the complete tree is sampled stagewise and therefore a *random tree* is produced.
The method is typically used for discrete, but very large trees with high bushiness, where only smaller "subtrees" are sampled. In this case, a lower bound estimate can be produced as in the two-stage case. For overall convergence, the bushyness of the sampled tree should increase to the bushiness of the large tree.