

Chapter 15

Approximations for Probability Distributions and Stochastic Optimization Problems

Georg Ch. Pflug and Alois Pichler

Abstract In this chapter, an overview of the scenario generation problem is given. After an introduction, the basic problem of measuring the distance between two single-period probability models is described in Section 15.2. Section 15.3 deals with finding good single-period scenarios based on the results of the first section. The distance concepts are extended to the multi-period situation in Section 15.4. Finally, Section 15.5 deals with the construction and reduction of scenario trees.

Keywords Scenario generation · Probability distances · Optimal discretizations · Scenario trees · Nested distributions

15.1 Introduction

Decision making under uncertainty is based upon

- (i) a probability model for the uncertain values,
- (ii) a cost or profit function depending on the decision variables and the uncertainties, and
- (iii) a probability functional, like expectation, median, etc., to summarize the random costs or profits in a real-valued objective.

We describe the decision model under uncertainty by

$$(Opt) \quad \max\{F(x) = \mathcal{A}_P[H(x, \xi)] : x \in \mathbb{X}\}, \quad (15.1)$$

where $H(\cdot, \cdot)$ denotes the profit function, with x the decision and ξ the random variable or random vector modeling uncertainty. Both – the uncertain data ξ and the decision x – may be stochastic processes adapted to some filtration \mathcal{F} . \mathcal{A} is the probability functional and \mathbb{X} is the set of constraints. P denotes the underlying probability measure.

G. Ch. Pflug (✉)

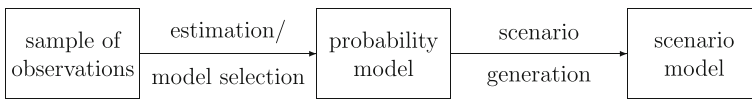
Department of Statistics and Operations Research, University of Vienna A-1010,

Wien – Vienna, Austria

e-mail: georg.pflug@univie.ac.at

The most difficult part in establishing a formalized decision model of type (15.1) for a real decision situation is to find the probability model P . Typically, there is a sample of past data available, but not more. It needs two steps to come from the sample of observations to the scenario model:

- (1) In the first step a *probability model* is identified, i.e., the description of the uncertainties as random variables or random processes by identifying the probability distribution. This step is based on statistical methods of model selection and parameter estimation. If several probability measures represent the data equally well, one speaks of *ambiguity*. In the non-ambiguous situation, one and only one probability model is selected and this model is the basis for the next step.
- (2) In the following scenario generation step, a *scenario model* is found, which is an approximation of (15.1) by a *finite model* of lower complexity than the probability model.



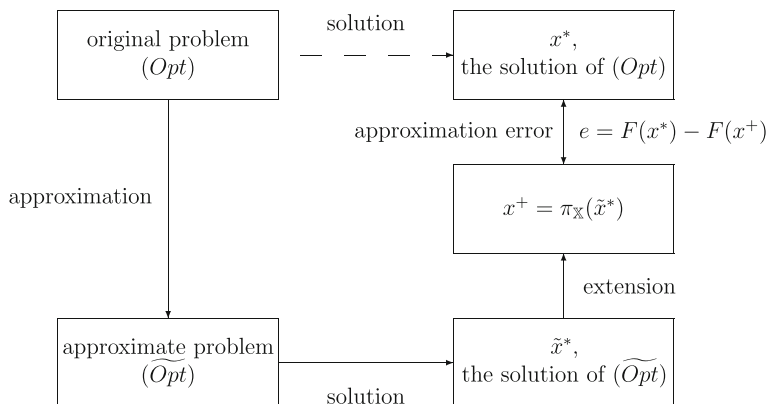
The scenario model differs from the original model (15.1) insofar as P is replaced by \tilde{P} , but – especially in multi-period problems – also the feasible set \mathbb{X} is replaced by a smaller, feasible set $\tilde{\mathbb{X}}$:

$$(\widetilde{Opt}) \quad \max\{\tilde{F}(\tilde{x}) = \mathcal{A}_{\tilde{P}}[H(\tilde{x}, \xi)] : \tilde{x} \in \tilde{\mathbb{X}}\}. \tag{15.2}$$

It is essential that the scenario model is finite (i.e., only finitely many values of the uncertainties are possible), but a good approximation to the original model as well. The finiteness is essential to keep the computational complexity low, and the approximative character is needed to allow the results to be carried over to the original model. In some cases, which do not need to be treated further here, finitely many scenarios are given right from the beginning (e.g., Switzerland will join the EU in 2015 – yes or no, or UK will adopt the EURO in 2020 and so on). The same is true, if the scenarios are just taken as the historic data without further transformation or information reduction. However, the justification of using just the empirical sample as the scenario model lies in the fact that these samples converge to the true underlying distribution if the sample size increases. For a quantification of this statement see Section 15.3.3 on Monte Carlo approximation below.

When stressing the approximative character of a scenario model we have to think about the consequences of this approximation and to estimate the error committed by replacing the probability model by the scenario model. Consider the diagram below: It is typically impossible to solve problem (15.1) directly due to its complexity. Therefore, one has to go around this problem and solve the simplified problem (15.2). Especially for multi-period problems, the solution of the approximate problem is not directly applicable to the original problem, but an extension function is needed, which extends the solution of the approximate problem to a solution of

the original problem. Of course, the extended solution of the approximate problem is typically suboptimal for the original problem. The respective gap is called the approximation error. It is the important goal of scenario generation to make this gap acceptably small.



The quality of a scenario approximation depends on the distance between the original probability P and the scenario probability \tilde{P} . Well-known quantitative stability theorems (see, e.g., Dupacova (1990), Heitsch et al. (2006), and Rachev and Roemisch (2002)) establish the relation between distances of probability models on one side and distances between optimal values or solutions on the other side (see, e.g., cite stability). For this reason, distance concepts for probability distributions on \mathbb{R}^M are crucial.

15.2 Distances Between Probability Distributions

Let \mathcal{P} be a set of probability measures on \mathbb{R}^M .

Definition 1.1 A *semi-distance* on \mathcal{P} is a function $d(\cdot, \cdot)$ on $\mathcal{P} \times \mathcal{P}$, which satisfies (i) and (ii) as follows:

- (i) **Nonnegativity:** For all $P_1, P_2 \in \mathcal{P}$

$$d(P_1, P_2) \geq 0.$$

- (ii) **Triangle inequality:** For all $P_1, P_2, P_3 \in \mathcal{P}$

$$d(P_1, P_2) \leq d(P_1, P_3) + d(P_3, P_2).$$

If a semi-distance satisfies the strictness property

- (iii) **Strictness:** If $d(P_1, P_2) = 0$, then $P_1 = P_2$, it is called a *distance*.

A general principle for defining semi-distances and distances consists in choosing a family of integrable functions \mathcal{H} (i.e., a family of functions such that the integral $\int h(w) dP(w)$ exists for all $P \in \mathcal{P}$) and defining

$$d_{\mathcal{H}}(P_1, P_2) = \sup \left\{ \left| \int h(w) dP_1(w) - \int h(w) dP_2(w) \right| : h \in \mathcal{H} \right\}.$$

We call $d_{\mathcal{H}}$ the (semi-)distance *generated by* \mathcal{H} .

In general, $d_{\mathcal{H}}$ is only a semi-distance. If \mathcal{H} is *separating*, i.e., if for every pair $P_1, P_2 \in \mathcal{P}$ there is a function $h \in \mathcal{H}$ such that $\int h dP_1 \neq \int h dP_2$, then $d_{\mathcal{H}}$ is a distance.

15.2.1 The Moment-Matching Semi-Distance

Let \mathcal{P}_q be the set of all probability measures on \mathbb{R}^1 which possess the q th moment, i.e., for which $\int \max(1, |w|^q) dP(w) < \infty$. The moment-matching semi-distance on \mathcal{P}_q is

$$d_{M_q}(P_1, P_2) = \sup \left\{ \left| \int w^s dP_1(w) - \int w^s dP_2(w) \right| : s \in \{1, 2, \dots, q\} \right\}. \tag{15.3}$$

The moment-matching semi-distance is not a distance, even if q is chosen to be large or even infinity. In fact, there are examples of two different probability measures on \mathbb{R}^1 , which have the same moments of all orders. For instance, there are probability measures, which have all moments equal to those of the lognormal distribution, but are not lognormal (see, e.g., Heyde (1963)).

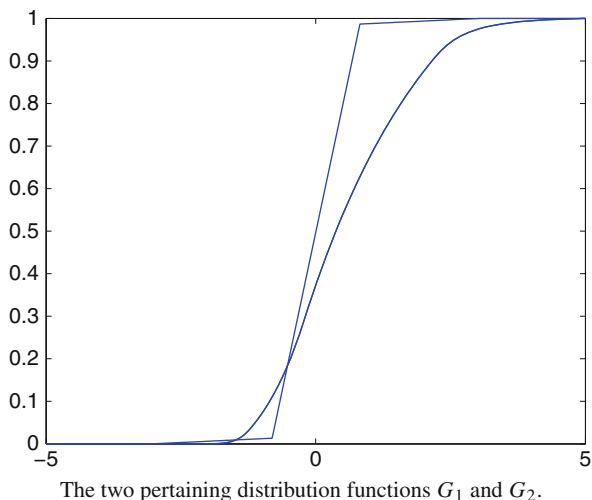
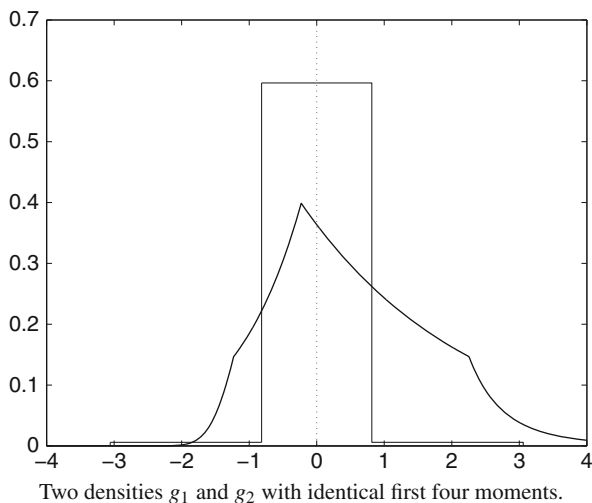
A widespread method is to match the first four moments (Hoyland and Wallace 2001), i.e., to work with d_{M_4} . The following example shows that two densities coinciding in their first four moments may exhibit very different properties.

Example Let P_1 and P_2 be the two probability measures with densities g_1 and g_2 where

$$\begin{aligned} g_1(x) &= 0.39876 [\exp(-|x + 0.2297|^3) \mathbb{1}_{\{x \leq -1.2297\}} \\ &\quad + \exp(-|x + 0.2297|) \mathbb{1}_{\{-1.2297 < x \leq -0.2297\}} \\ &\quad + \exp(-0.4024 \cdot (x + 0.2297)) \mathbb{1}_{\{-0.2297 < x \leq 2.2552\}} \\ &\quad + 1.09848 \cdot (0.4024x + 0.29245)^{-6} \mathbb{1}_{\{2.2552 < x\}}], \\ g_2(x) &= 0.5962 \mathbb{1}_{\{|x| \leq 0.81628\}} + 0.005948 \mathbb{1}_{\{0.81628 < |x| \leq 3.05876\}}. \end{aligned}$$

Both densities are unimodal and coincide in the first four moments, which are $m_1 = 0, m_2 = 0.3275, m_3 = 0, m_4 = 0.72299$ ($m_q(P) = \int w^q dP(w)$). The fifth moments, however, could not differ more: While P_1 has infinite fifth moment, the fifth moment of P_2 is zero. Density g_1 is asymmetric, has a sharp cusp at the

point -0.2297 and unbounded support. In contrast, g_2 is symmetric around 0, has a flat density there, has finite support, and possesses all moments. The distribution functions and quantiles differ drastically: We have that $G_{P_1}(0.81) = 0.6257$ and $G_{P_1}(-0.81) = 0.1098$, while $G_{P_2}(0.81) = 0.9807$ and $G_{P_2}(-0.81) = 0.0133$. Thus the probability of the interval $[-0.81, 0.81]$ is only 51% under P_1 , while it is 96% under P_2 .



Summarizing, the moment matching semi-distance is not well suited for scenario approximation, since it is not fine enough to capture the relevant quality of approximation.

The other extreme would be to choose as the generating class \mathcal{H} all measurable functions h such that $|h| \leq 1$. This class generates a distance, which is called the variational distance (more precisely twice the variational distance). It is easy to see that if P_1 resp. P_2 have densities g_1 resp. g_2 w.r.t. λ , then

$$\begin{aligned} & \sup \left\{ \int h dP_1 - \int h dP_2 : |h| \leq 1, h \text{ measurable} \right\} \\ &= \int |g_1(x) - g_2(x)| d\lambda(x) \\ &= 2 \sup\{|P_1(A) - P_2(A)| : A \text{ measurable set}\}. \end{aligned}$$

We call

$$d_V(P_1, P_2) = \sup\{|P_1(A) - P_2(A)| : A \text{ measurable set}\} \tag{15.4}$$

the *variational distance* between P_1 and P_2 .

The variational distance is a very fine distance, too fine for our applications: If P_1 has a density and P_2 sits on at most countably many points, then $d_V(P_1, P_2) = 1$, independently on the number of mass points of P_2 . Thus there is no hope to approximate any continuous distribution by a discrete one w.r.t. the variational distance.

One may, however, restrict the class of sets in (15.4) to a certain subclass. If one takes the class of half-unbounded rectangles in \mathbb{R}^M of the form $(-\infty, t_1] \times (-\infty, t_2] \times \dots \times (-\infty, t_M]$ one gets the *uniform distance*, also called *Kolmogorov distance*

$$d_U(P_1, P_2) = \sup\{|G_{P_1}(x) - G_{P_2}(x)| : x \in \mathbb{R}^M\}, \tag{15.5}$$

where $G_P(t)$ is the distribution function of P ,

$$G_P(t) = P\{(-\infty, t_1] \times \dots \times (-\infty, t_M]\}.$$

The uniform distance appears in the Hlawka–Koksma inequality:

$$\int h(u) dP_1(u) - \int h(u) dP_2(u) \leq d_U(P_1, P_2) \cdot V(h), \tag{15.6}$$

where $V(h)$ is the Hardy–Krause variation of h . In dimension $M = 1$, V is the usual total variation

$$V(h) = \sup \left\{ \sum_i |h(z_i) - h(z_{i-1})| : z_1 < z_2 < \dots < z_n, n \text{ arbitrary} \right\}.$$

In higher dimensions, let $V^{(M)}(h) = \sup \sum_{J_1, \dots, J_n}$ is a partition by rectangles J_i $|\Delta_{J_i}(h)|$, where $\Delta_J(h)$ is the sum of values of h at the vertices of J , where adjacent vertices get

opposing signs. The Hardy–Krause variation of h is defined as $\sum_{m=1}^M V^{(m)}(h)$. d_U is invariant w.r.t. monotone transformations, i.e., if T is a monotone transformation, then the image measures $P_i^T = P_i \circ T^{-1}$ satisfy

$$d_U(P_1^T, P_2^T) = d_U(P_1, P_2).$$

Notice that a unit mass at point 0 and at point $1/n$ are at a distance 1 both in the d_U distance, for any even large n . Thus it may be felt that also this distance is too fine for good approximation results. A minimal requirement of closedness is that integrals of bounded, continuous functions are close and this leads to the notion of weak convergence.

Definition 1.2 (Weak convergence) A sequence of probability measures P_n on \mathbb{R}^M converges weakly to P (in symbol $P_n \Rightarrow P$), if

$$\int h dP_n \rightarrow \int h dP$$

as $n \rightarrow \infty$, for all bounded, continuous functions h . Weak convergence is the most important notion for scenario approximation: We typically aim at constructing scenario models which, for increasing complexity, finally converge weakly to the underlying continuous model.

If a distance d has the property that

$$d(P_n, P) \rightarrow 0 \quad \text{if and only if } P_n \Rightarrow P$$

we say that d *metricizes weak convergence*.

The following three distances metricize the weak convergence on some subsets of all probability measures on \mathbb{R}^M :

- (i) *The bounded Lipschitz metric:* If \mathcal{H} is the class of bounded functions with Lipschitz constant 1, the following distance is obtained:

$$d_{BL}(P_1, P_2) = \sup \left\{ \int h dP_1 - \int h dP_2 : |h(u)| \leq 1, L_1(h) \leq 1 \right\}, \quad (15.7)$$

where

$$L_1(h) = \inf\{L : |h(u) - h(v)| \leq L\|u - v\|\}.$$

This distance metricizes weak convergence for all probability measures on \mathbb{R}^M .

- (ii) *The Kantorovich distance:* If one drops the requirement of boundedness of h , one gets the Kantorovich distance

$$d_{KA}(P_1, P_2) = \sup \left\{ \int h dP_1 - \int h dP_2 : L_1(h) \leq 1 \right\}. \tag{15.8}$$

This distance metricizes weak convergence on sets of probability measures which possess uniformly a first moment. (A set \mathcal{P} of probability measures has uniformly a first moment, if for every ϵ there is a constant C_ϵ such that $\int_{\|x\| > C_\epsilon} \|x\| dP(x) < \epsilon$ for all $P \in \mathcal{P}$.) See the review of Gibbs and Su (2002). On the real line, the Kantorovich metric may also be written as

$$d_{KA}(P, \tilde{P}) = \int |G_P(u) - G_{\tilde{P}}(u)| du = \int |G_P^{-1}(u) - G_{\tilde{P}}^{-1}(u)| du,$$

where $G_P^{-1}(u) = \sup\{v : G_P(v) \leq u\}$ (see Vallander (1973)).

(iii) *The Fortet–Mourier metric:* If \mathcal{H} is the class of Lipschitz functions of order q , the Fortet–Mourier distance is obtained:

$$d_{FM_q}(P_1, P_2) = \sup \left\{ \int h dP_1 - \int h dP_2 : L_q(h) \leq 1 \right\}, \tag{15.9}$$

where the Lipschitz constant of order q is defined as

$$L_q(h) = \inf\{L : |h(u) - h(v)| \leq L\|u - v\| \max(1, \|u\|^{q-1}, \|v\|^{q-1})\}. \tag{15.10}$$

Notice that $L_q(h) \leq L_{q'}(h)$ for $q' \leq q$. In particular, $L_q(h) \leq L_1(h)$ for all $q \geq 1$ and therefore

$$d_{FM_q}(P_1, P_2) \geq d_{FM_{q'}}(P_1, P_2) \geq d_{KA}(P_1, P_2).$$

The Fortet–Mourier distance metricizes weak convergence on sets of probability measures possessing uniformly a q th moment. Notice that the function $u \mapsto \|u\|^q$ is q -Lipschitz with Lipschitz constant $L_q = q$. On \mathbb{R}^1 , the Fortet–Mourier distance may be equivalently written as

$$d_{FM_q}(P_1, P_2) = \int \max(1, |u|^{q-1}) |G_{P_1}(u) - G_{P_2}(u)| du$$

(see Rachev (1991), p. 93). If $q = 1$, the Fortet–Mourier distance coincides with the Kantorovich distance.

The Kantorovich distance is related to the mass transportation problem (Monge’s problem – Monge (1781), see also Rachev (1991), p. 89) through the following theorem.

Theorem 1.3 (Kantorovich–Rubinstein)

$$\begin{aligned} d_{\text{KA}}(P_1, P_2) = \inf\{\mathbb{E}(|X_1 - X_2|) : \text{s.t. the joint distribution } (X_1, X_2) \\ \text{is arbitrary, but the marginal distributions are fixed} \\ \text{such that } X_1 \sim P_1, X_2 \sim P_2\} \end{aligned} \quad (15.11)$$

(see Rachev (1991), theorems 5.3.2 and 6.1.1).

The infimum in (15.11) is attained. The optimal joint distribution (X_1, X_2) describes how masses with distribution P should be transported with minimal effort to yield the new mass distribution \tilde{P} . The analogue of the Hlawka–Koksma inequality is here

$$\int |h(u) dP_1(u) - \int h(u) dP_2(u)| \leq d_{\text{KA}}(P_1, P_2) \cdot L_1(h). \quad (15.12)$$

The Kantorovich metric can be defined on arbitrary metric spaces R with metric d : If P_1 and P_2 are probabilities on this space, then $d_{\text{KA}}(P_1, P_2; d)$ is defined by

$$d_{\text{KA}}(P_1, P_2; d) = \sup \left\{ \int h dP_1 - \int h dP_2 : |h(u) - h(v)| \leq d(u, v) \right\}. \quad (15.13)$$

Notice that the metric $d(\cdot, \cdot; d)$ is compatible with the metric d in the following sense: If δ_u resp. δ_v are unit point masses at u resp. v , then

$$d_{\text{KA}}(\delta_u, \delta_v; d) = d(u, v).$$

The theorem of Kantorovich–Rubinstein extends to the general case:

$$\begin{aligned} d_{\text{KA}}(P_1, P_2; d) = \inf\{\mathbb{E}[d(X_1, X_2)] : \text{where the joint distribution } (X_1, X_2) \\ \text{is arbitrary, but the marginal distributions are fixed} \\ \text{such that } X_1 \sim P_1, X_2 \sim P_2\}. \end{aligned} \quad (15.14)$$

A variety of Kantorovich metrics may be defined, if \mathbb{R}^M is endowed with different metrics than the Euclidean one.

15.2.2 Alternative Metrics on \mathbb{R}^M

By (15.13), to every metric d on \mathbb{R}^M , there corresponds a distance of the Kantorovich type. For instance, let d_0 be the discrete metric

$$d_0(u, v) = \begin{cases} 1 & \text{if } u \neq v, \\ 0 & \text{if } u = v. \end{cases}$$

The set of all Lipschitz functions w.r.t. the discrete metric coincides with the set of all measurable functions h such that $0 \leq h \leq 1$ or its translates. Consequently the pertaining Kantorovich distance coincides with the variational distance

$$d_{KA}(P_1, P_2; d_0) = d_V(P_1, P_2).$$

Alternative metrics on \mathbb{R}^1 can, for instance, be generated by nonlinear transform of the axis. Let χ be any bijective monotone transformation, which maps \mathbb{R}^1 into \mathbb{R}^1 . Then $d(u, v) := |\chi(u) - \chi(v)|$ defines a new metric on \mathbb{R}^1 . Notice that the family functions which are Lipschitz w.r.t the distance d and the Euclidean distance $|u - v|$ may be quite different.

In particular, let us look at the bijective transformations (for $q > 0$)

$$\chi_q(u) = \begin{cases} u & |u| \leq 1 \\ |u|^q \operatorname{sgn}(u) & |u| \geq 1 \end{cases}.$$

Notice that $\chi_q^{-1}(u) = \chi_{1/q}(u)$. Introduce the metric $d_{\chi_q}(u, v) = |\chi_q(u) - \chi_q(v)|$. We remark that $d_{\chi_1}(u, v) = |u - v|$. Denote by $d_{KA}(\cdot, \cdot; d_{\chi_q})$ the Kantorovich distance which is based on the distance d_{χ_q} ,

$$d_{KA}(P_1, P_2; d_{\chi_q}) = \sup \left\{ \int h dP_1 - \int h dP_2 : \|h(u) - h(v)\| \leq d_{\chi_q}(u, v) \right\}.$$

Let P^{χ_q} be the image measure of P under χ_q , that is P^{χ_q} assigns to a set A the value $P\{\chi_{1/q}(A)\}$. Notice that P^{χ_q} has distribution function

$$G_{P^{\chi_q}}(x) = G_P(\chi_{1/q}(x)).$$

This leads to the following identity:

$$d_{KA}(P_1, P_2; d_{\chi_q}) = d_{KA}(P_1^{\chi_q}, P_2^{\chi_q}; |\cdot|).$$

It is possible to relate the Fortet–Mourier distance d_{M_q} to the distance $d_{KA}(\cdot, \cdot; d_{\chi_q})$. To this end, we show first that

$$L_q(h \circ \chi_q) \leq q \cdot L_1(h) \tag{15.15}$$

and

$$L_1(h \circ \chi_{1/q}) \leq L_q(h). \tag{15.16}$$

If $L_1(h) < \infty$, then

$$\begin{aligned} |h(\chi_q(u)) - h(\chi_q(v))| &\leq L_1(h) |\chi_q(u) - \chi_q(v)| \\ &\leq L_1(h) \cdot q \cdot \max(1, |u|^{q-1}, |v|^{q-1}) |u - v| \end{aligned}$$

which implies (15.15). On the other hand, if $L_q(h) < \infty$, then

$$\begin{aligned} & |h(\chi_{1/q}(u)) - h(\chi_{1/q}(v))| \\ & \leq L_q(h) \max(1, |\chi_{1/q}(u)|^{q-1}, |\chi_{1/q}(v)|^{q-1}) |\chi_{1/q}(u) - \chi_{1/q}(v)| \\ & \leq L_q(h) \max(1, \max(|u|, |v|)^{q-1}) \max(1, \max(|u|, |v|)^{(1-q)/q}) |u - v| \\ & \leq L_q(h) |u - v| \end{aligned}$$

and therefore (15.16) holds. As a consequence, we get the following relations:

$$\begin{aligned} \frac{1}{q} \mathbf{d}_{\text{KA}}(P_1, P_2; d_{\chi_q}) &= \frac{1}{q} \mathbf{d}_{\text{KA}}(G_{P_1} \circ \chi_{1/q}, G_{P_2} \circ \chi_{1/q}) \\ &\leq \mathbf{d}_{\text{FM}_q}(P_1, P_2) \\ &\leq \mathbf{d}_{\text{KA}}(G_{P_1} \circ \chi_{1/q}, G_{P_2} \circ \chi_{1/q}) = \mathbf{d}_{\text{KA}}(P_1, P_2; d_{\chi_q}). \end{aligned} \quad (15.17)$$

Thus one sees that it is practically equivalent to measure the distance by the Fortet–Mourier distance of order q or by the Kantorovich distance (i.e., the Fortet–Mourier distance of order 1) but with the real line endowed with the metric d_{χ_q} .

15.2.3 Extending the Kantorovich Distance to the Wasserstein Distance

By the Kantorovich–Rubinstein Theorem, the Kantorovich metric is the optimal cost of a transportation problem. If the cost function is not the distance d itself, but a power d^r of it, we are led to the *Wasserstein metric* $\mathbf{d}_{\text{W},r}$.

$$\mathbf{d}_{\text{W},r}(P_1, P_2; d)^r = \inf \left\{ \int d(x_1, x_2)^r \pi(dx_1, dx_2); \text{ s.t. the joint probability measure } \pi \text{ has marginals such that } \pi(A \times \mathbb{R}^M) = P_1(A) \text{ and } \pi(\mathbb{R}^M \times B) = P_2(B) \right\}. \quad (15.18)$$

This is a metric for $r \geq 1$. For $r < 1$, the Wasserstein distance is defined as

$$\mathbf{d}_{\text{W},r}(P_1, P_2; d) = \inf \left\{ \int d(x_1, x_2)^r \pi(dx_1, dx_2); \text{ where the joint probability measure } \pi \text{ has marginals such that } \pi(A \times \mathbb{R}^M) = P_1(A) \text{ and } \pi(\mathbb{R}^M \times B) = P_2(B) \right\} \quad (15.19)$$

(see Villani (2008)).

15.2.4 Notational Convention

When there is no doubt about the underlying distance d , we simply write

$$d_r(P_1, P_2) \quad \text{instead of} \quad d_{W,r}(P_1, P_2; d).$$

Recall that in this notation, the Kantorovich distance is $d_{KA}(P_1, P_2) = d_1(P_1, P_2)$.

15.2.5 Wasserstein Distance and Kantorovich Distance as Linear Program

To compute the Wasserstein distance for discrete probability distributions amounts to solving a linear program, as we shall outline here:

Consider the discrete measures $P := \sum_s p_s \delta_{z_s}$ and $Q := \sum_t q_t \delta_{z'_t}$ and – within this setting – the problem

$$\begin{aligned} & \text{Minimize} && \sum_{s,t} \pi_{s,t} c_{s,t} \\ & \text{(in } \pi) && \\ & \text{subject to} && \sum_t \pi_{s,t} = p_s, \\ & && \sum_s \pi_{s,t} = q_t \text{ and} \\ & && \pi_{s,t} \geq 0, \end{aligned} \tag{15.20}$$

where $c_{s,t} := d(z_s, z'_t)^r$ is a matrix derived from the general distance function d . This matrix c represents the weighted costs related to the transportation of masses from z_s to z'_t .

Notice that (15.20) is a general linear program, which is already well defined by specifying the cost matrix $c_{s,t}$ and the probability masses p_s and q_t . So in particular the mass points z_s are not essential to formulate the problem; they may represent abstract objects.

Observe further that

$$\sum_{s,t} \pi_{s,t} = \sum_s p_s = \sum_t q_t = 1,$$

thus π is a probability measure, representing – due to the constraints – a transport plan from P to Q . The linear program (15.20) thus returns the minimal distance between those measures, and as a minimizing argument the optimal transport plan in particular for transporting masses from object z_s to z'_t .

The linear program (15.20) has a dual with vanishing duality gap. The theorem of Kantorovich–Rubinstein allows to further characterize (in the situation $r = 1$) the dual program; it amounts to just finding a Lipschitz-1 function h which maximizes the respective expectation:

$$\begin{aligned} & \text{Maximize} && \sum_s p_s h_s - \sum_t q_t h_t, \\ & \text{(in } h) && \\ & \text{subject to} && h_s - h_t \leq d(z_s, z_t). \end{aligned}$$

Due to the vanishing duality gap the relation

$$\sum_s p_s h_s^* - \sum_t q_t h_t^* = \sum_{s,t} \pi_{s,t}^* c_{s,t}$$

holds true, where h^* is the optimal Lipschitz-1 function and π^* the optimal joint measure.

15.2.6 Bibliographical Remarks

The distance d_{KA} was introduced by Kantorovich in 1942 as a distance in general spaces. In 1948, he established the relation of this distance (in \mathbb{R}^n) to the mass transportation problem formulated by Gaspard Monge in 1781. In 1969, L. N. Wasserstein – unaware of the work of Kantorovich – reinvented this distance for using it for convergence results of Markov processes and 1 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. The distance d_U was introduced by Kolmogorov in 1933. It is often called Kolmogorov distance.

15.3 Single-Period Discretizations

In this section, we study the *discrete approximation problem*:

Let a probability measure P on \mathbb{R}^M be given. We want to find a probability measure sitting on at most S points, such that some distance $d(P, \tilde{P})$ is small.

We say “small” and not “minimal”, since it may be computationally intractable to really find the minimum. Formally, define \mathcal{P}_S as the family of all probability measures

$$\tilde{P} = \sum_{s=1}^S p_s \delta_{z_s}$$

on \mathbb{R}^M sitting on at most S points. Here δ_z denotes the point mass at point z . We try to find a $\tilde{P} \in \mathcal{P}_S$ such that

$$d(P, \tilde{P}) \text{ is close to } \min\{d(P, Q) : Q \in \mathcal{P}_S\}.$$

There are several methods to attack the problem, which we order in decreasing order of computational effort and as well in decreasing order of approximation quality (see Fig. 15.1):

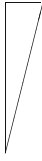

	Approximation quality	Computational ease
Optimal quantization	 high	 low
Quantization heuristics		
Quasi-Monte Carlo		
Monte Carlo	low	high

Fig. 15.1 Methods of scenario generation

While the probability measure P may have a density or may be discrete, the measure \tilde{P} is always characterized by the list of probability masses $p_s, s = 1, \dots, S$ and the list of mass points $z_s, s = 1, \dots, S$. The discrete measure \tilde{P} is represented by

$$\tilde{P} = \begin{bmatrix} p_1 & p_2 & \cdots & p_S \\ z_1 & z_2 & \cdots & z_S \end{bmatrix}.$$

The approximation problem depends heavily on the chosen distance d , as will be seen. We treat the optimal quantization problem in Section 15.3.1, the quasi-Monte Carlo technique in Section 15.3.2 and the Monte Carlo technique in Section 15.3.3. Some quantization heuristics are contained in Chapter 3.

15.3.1 Optimal and Asymptotically Optimal Quantizers

The basic approximation problem, i.e., to approximate a distribution on \mathbb{R}^M by a distribution sitting on at most S points leads to a non-convex optimization problem. To find its global solution is computationally hard. Moreover, no closed-form solution is known, even for the simplest distributions.

Recall the basic problem: Let d be some distance for probability measures and let \mathcal{P}_S be the family of probability distributions sitting on at most S points. For a given probability $P \in \mathcal{P}_S$, the main question is to find the *quantization error*

$$q_{S,d}(P) = \inf\{d(P, Q) : Q \in \mathcal{P}_S\} \tag{15.21}$$

and the *optimal quantization set*

$$\mathcal{Q}_{S,d}(P) = \operatorname{argmin} \{d(P, Q) : Q \in \mathcal{P}_S\} \tag{15.22}$$

(if it exists).

15.3.1.1 Optimal Quantizers for Wasserstein Distances d_r

Since all elements \tilde{P} of \mathcal{P}_S are of the form $\tilde{P} = \sum_{s=1}^S p_s \delta_{z_s}$, the quantization problem consists essentially of two steps:

- (1) Find the optimal supporting points z_s .
- (2) Given the supporting points z_s , find the probabilities p_s , which minimize

$$d_r \left(P, \sum_{s=1}^S p_s \delta_{z_s} \right) \quad (15.23)$$

for the Wasserstein distance.

Recall the Wasserstein d_r distances contain the Kantorovich distance as the special case for $r = 1$.

We show that the second problem has an easy solution for the chosen distance: Let $z = (z_1, \dots, z_S)$ be the vector of supporting points and suppose that they are all distinct. Introduce the family of *Voronoi diagrams* $\mathcal{V}(z)$ as the family of all partitions (A_1, \dots, A_S) , where A_s are pairwise disjoint and $\mathbb{R}^M = \bigcup_{s=1}^S A_s$, such that

$$A_s \subseteq \{y : d(y, z_s) = \min_k d(y, z_k)\}.$$

Then the possible optimal probability weights p_s for minimizing $d_r \left(P, \sum_{s=1}^S p_s \delta_{z_s} \right)$ can be found by

$$p = (p_1, \dots, p_S), \text{ where } p_s = P(A_s); \quad (A_1, \dots, A_S) \in \mathcal{V}(z).$$

The optimal probability weights are unique, iff

$$P\{y : d(y, z_s) = d(y, z_k), \text{ for some } s \neq k\} = 0.$$

One may also reformulate (15.23) in terms of random variables: Let $X \sim P$ and let \tilde{X} be a random variable which takes only finitely many values $\tilde{X} \in \{z_1, \dots, z_S\}$, then

$$\begin{aligned} D_{d_r}(z) &:= \inf\{\mathbb{E}[d(X, \tilde{X})^r] : X \sim P, \tilde{X} \in \{z_1, \dots, z_S\}\} \\ &= \inf\{d_r(P, \tilde{P})^r : \text{supp}(\tilde{P}) \subseteq \{z_1, \dots, z_S\}\} \\ &= \int \min_s d(x, z_s)^r dP(x). \end{aligned}$$

Unfortunately, the function $z \mapsto D_{d_r}(z)$ is non-convex and typically has multiple local minima. Notice that the quantization error defined in (15.21) satisfies

$$q_{S, d_r}(P) = \min\{[D_{d_r}(z)]^{1/r} : z = (z_1, \dots, z_S)\}.$$

Lemma 2.1 The mapping $P \mapsto q_{S,d_r}(P)$ is concave.

Proof Let $P = \sum_{i=1}^I w_i P_i$ for some positive weights with $\sum_{i=1}^I w_i = 1$. Then

$$q_{S,d_r}(P) = \sum_{i=1}^I w_i \int \min[d(x, z_s)]^r dP_i(x) \geq \sum_{i=1}^I w_i q_{S,d_r}(P_i).$$

□

Lemma 2.2 Let $\sum_{i=1}^I S_i \leq S$ and the weights w_i as above. Then

$$q_{S,d_r} \left(\sum_{i=1}^I w_i P_i \right) \leq \sum_{i=1}^I w_i q_{S_i,d_r}(P_i).$$

Proof Let \tilde{P}_i such that $d_r(P_i, \tilde{P}_i) = q_{S_i,d_r}(P_i)$. Then

$$q_{S,d_r} \left(\sum_{i=1}^I w_i P_i \right) \leq d_r \left(\sum_{i=1}^I w_i P_i, \sum_{i=1}^I w_i \tilde{P}_i \right) \leq \sum_{i=1}^I w_i q_{S_i,d_r}(P_i).$$

□

For the next lemma, it is necessary to specify the distance d to

$$d_r(x, y) = \left[\sum_{m=1}^M |x_m^r - y_m^r| \right]^{1/r}.$$

Lemma (Product quantizers) Let $\prod_{i=1}^I S_i \leq S$. Then

$$q_{S,d_r}(\otimes_{i=1}^I P_i) \leq \sum_{i=1}^I q_{S_i,d_r}(P_i).$$

Proof Let $\tilde{P}_i \in \mathcal{P}_{S_i}$ such that $d_r(P_i, \tilde{P}_i) = q_{S_i,d_r}(P_i)$. Then

$$q_{S,d_r}(\otimes_{i=1}^I P_i) \leq d_r(\otimes_{i=1}^I P_i, \otimes_{i=1}^I \tilde{P}_i) = \sum_{i=1}^I q_{S_i,d_r}(P_i).$$

□

15.3.1.2 Interpretation as Facility Location Problems on \mathbb{R}

On \mathbb{R}^1 one may w.l.o.g. assume that the points z_s are in ascending order $z_1 \leq z_2 \leq \dots \leq z_S$. The pertaining Voronoi tessellation then is

$$V_s = (b_{s-1}, b_s], \quad s = 1, \dots, S,$$

with $b_0 = -\infty$, $b_S = +\infty$ and for $1 \leq s \leq S$; the points b_s are chosen such that $d(b_s, z_s) = d(b_s, z_{s+1})$, and the optimal masses are then $p_s = G(b_s) - G(b_{s-1})$. For the Euclidean distance on \mathbb{R} , the b_s 's are just the midpoints

$$z_s = \frac{1}{2}(b_{s-1} + b_s), \quad s = 1, \dots, S.$$

In this case, the Wasserstein distance is

$$d_r(P, \tilde{P})^r = \sum_{s=1}^S \int_{\frac{z_{s-1}+z_s}{2}}^{\frac{z_s+z_{s+1}}{2}} |u - z_s|^r dG(u), \quad (15.24)$$

with $z_0 = -\infty$ and $z_{S+1} = +\infty$.

While step (2) of (15.23) is easy, step (1) is more involved: Let, for $z_1 \leq z_2 \leq \dots \leq z_S$,

$$D(z_1, \dots, z_S) = \int \min_s d(x, z_s)^r dG(x) \quad (15.25)$$

The global minimum of this function is sought for. Notice that D is non-convex, i.e., the global minimization of D is a hard problem. The function D can be viewed as the travel mean costs of customers distributed along the real line with d.f. G to travel to the nearest location of one of the facilities, placed at z_1, \dots, z_S , if the travel costs from u to v are $d(u, v)^r$.

Facility Location The problem of minimizing the function D in (15.25) is an instance of the classical Weber location problem, which was introduced by A. Weber in 1909 (English translation published 1929 (Weber 1929)). A recent state-of-the-art overview of the theory and applications was published by Drezner and Hamacher (2002).

Explicit Optimal Solutions Only in exceptional cases explicit solutions of the optimal quantization problems (15.21) resp. (15.22) are known. Here are some examples; for more details, see Graf and Luschgy (2000).

Example (The Laplace distribution in \mathbb{R}^1) For the Euclidean distance on \mathbb{R} and the Laplace distribution with density $f(x) = \frac{1}{2} \exp(-|x|)$, the optimal supporting points for even $S = 2k$ are

$$z_s = \begin{cases} 2 \log \left(\frac{s}{\sqrt{k^2+k}} \right), & 1 \leq s \leq k, \\ 2 \log \left(\frac{\sqrt{k^2+k}}{S+1-s} \right), & k+1 \leq s \leq S. \end{cases}$$

The quantization error is

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

Example (The Exponential distribution in \mathbb{R}^1) For the Euclidean distance on \mathbb{R} and the Exponential distribution with density $f(x) = \exp(-x)\mathbb{1}_{x \geq 0}$, the optimal supporting points are

$$z_s = 2 \log\left(\frac{\sqrt{S^2 + S}}{S + 1 - s}\right); \quad s = 1, \dots, S.$$

The quantization error is

$$q_{S,d_1}(P) = \log\left(1 + \frac{1}{S}\right).$$

Example (The uniform distribution in \mathbb{R}^M) To discuss the uniform distribution in higher dimension, it is necessary to further specify the distance on the sample space:

As distances on \mathbb{R}^M we consider the p -norms $d_p(u, v) = \left[\sum_{m=1}^M |u_m - v_m|^p\right]^{1/p}$ for $1 \leq p < \infty$ resp. $d_\infty(u, v) = \max_{1 \leq m \leq M} |u_m - v_m|$.

Based on d_p , we consider the Wasserstein distances $d_{p,r}$

$$d_{p,r}(P_1, P_2)^r = \inf \left\{ \int d_p(x_1, x_2)^r \pi(dx_1, dx_2); \text{ where the joint probability measure } \pi \text{ has marginals such that} \right. \\ \left. \pi(A \times \mathbb{R}^M) = P_1(A) \text{ and } \pi(\mathbb{R}^M \times B) = P_2(B) \right\}. \quad (15.26)$$

Let $q_{S,d_{p,r}}^{(M)} := q_{S,d_{p,r}}(\mathcal{U}[0, 1]^M)$ be the minimal error of approximation to the uniform distribution on the M -dimensional unit cube $[0, 1]^M$. The exact values are only known for

$$q_{1,d_{p,r}}^{(M)} = \begin{cases} \frac{M}{(1+r)2^r}, & 1 \leq p < \infty, \\ \frac{M}{(M+r)2^r}, & p = \infty \end{cases}$$

and

$$q_{1,d_{\infty,r}}^{(M)} = \frac{M}{(M+r)2^r}.$$

15.3.1.3 Asymptotically Optimal Quantizers

In most cases, the solutions of (15.21) and (15.22) are unknown and cannot be expressed in an analytic way. Therefore one may ask the simpler ‘‘asymptotic’’ questions:

- What is the rate in which $q_{S,d_r}(P)$ converges to zero as S tends to infinity?
- Is there a constructive way to find a sequence being asymptotically optimal, i.e., a sequence (P_S^+) such that

$$\frac{d(P, P_S^+)}{q_{S,d_r}(P)} \rightarrow 1$$

as $S \rightarrow \infty$?

To investigate these questions introduce the quantities

$$\begin{aligned} \bar{q}_{d_{p,r}}^{(M)} &:= \inf_S S^{r/M} q_{S,d_{p,r}}(\mathcal{U}[0, 1]^M), \\ \bar{q}_{d_{p,r}}(P) &:= \inf_S S^{r/M} q_{S,d_{p,r}}(P), \end{aligned} \tag{15.27}$$

where $\mathcal{U}[0, 1]^M$ is the uniform distribution on the M -dimensional unit cube $[0, 1]^M$.

It deduces from the concavity of $P \mapsto q_{S,d_{p,r}}(P)$ that the mappings $P \mapsto \bar{q}_{d_{p,r}}(P)$ are concave as well.

The quantities $\bar{q}_{d_{p,r}}^{(M)}$ are found to be universal constants; however, they are known in special cases only:

- $\bar{q}_{d_{p,r}}^{(1)} = \frac{1}{(1+r)2^r}$,
- $\bar{q}_{d_{\infty,r}}^{(M)} = \frac{M}{(M+r)2^r}$.

The next theorem links a general distribution P with the uniform distribution, providing a relation between $\bar{q}_{d_{p,r}}(P)$ and $\bar{q}_{d_{p,r}}^{(M)}$. As for a proof we refer the reader to Graf and Luschgy (2000), theorem 6.2. or Na and Neuhoff (1995).

Theorem (Zador–Gersho formula) *Suppose that P has a density g with $\int |u|^{r+\delta} g(u) du < \infty$ for some $\delta > 0$. Then*

$$\bar{q}_{d_{p,r}}(P) = \inf_S S^{r/M} q_{S,d_{p,r}}(P) = \bar{q}_{d_{p,r}}^{(M)} \cdot \left[\int_{\mathbb{R}^M} g(x)^{\frac{M}{M+r}} dx \right]^{\frac{M+r}{M}}, \tag{15.28}$$

where $\bar{q}_{d_{p,r}}^{(M)}$ is given by (15.27).

For the real line, this specializes to

$$\bar{q}_{d_{p,r}}(P) = \frac{1}{(1+r)2^r} \left(\int_{\mathbb{R}} g(u)^{\frac{1}{1+r}} du \right)^{1+r}$$

for all p , since all distances d_p are equal for $M = 1$. Specializing further to $r = 1$, one gets

$$\bar{q}_{d_{p,1}}(P) = \frac{1}{4} \left(\int_{\mathbb{R}} \sqrt{g(u)} \, du \right)^2.$$

Heuristics The Zador–Gersho formula (15.28), together with identity (15.24) give rise that the optimal, asymptotic point density for z_s is proportional to $g^{\frac{M}{M+r}}$. In \mathbb{R}^1 again this translates to solving the quantile equations

$$\int_{-\infty}^{z_s} g^{\frac{1}{1+r}}(x) \, dx = \frac{2s-1}{2S} \cdot \int_{-\infty}^{\infty} g^{\frac{1}{1+r}}(x) \, dx$$

for $s = 1, 2, \dots, S$ and then to put

$$p_s := \int_{\frac{z_{s-1} + z_s}{2}}^{\frac{z_s + z_{s+1}}{2}} g(x) \, dx$$

as above. It can be proved that the points z_s specified accordingly are indeed asymptotically optimal and in practice this has proven to be a very powerful setting.

Example If P is the univariate normal distribution $N(\mu, \sigma^2)$, then $\bar{q}_{d_{p,1}}(N(\mu, \sigma) = \sigma \sqrt{\frac{\pi}{2}}$. For the M -dimensional multivariate normal distribution $N(\mu, \Sigma)$ one finds that

$$\bar{q}_{d_{p,r}}(N(\mu, \Sigma)) = \bar{q}_{d_{p,r}}^{(M)} (2\pi)^{r/2} \left(\frac{M+r}{M} \right)^{(M+r)/2} (\det \Sigma)^{r/(2M)}.$$

15.3.1.4 Optimal Quantizers for the Uniform (Kolmogorov) Distance

In dimension $M = 1$, the solution of the uniform discrete approximation problem is easy: Notice that the uniform distance is invariant w.r.t. monotone transformations, which means that the problem can be mapped to the uniform distribution via the quantile transform and then mapped back.

The uniform distribution $\mathcal{U}[0, 1]$ is optimally approximated by

$$\tilde{P} = \left[\begin{array}{cccc} 1/S & 1/S & 1/S & \cdots & 1/S \\ \frac{1}{2S} & \frac{3}{2S} & \frac{5}{2S} & \cdots & \frac{2S-1}{2S} \end{array} \right]$$

and the distance is

$$d_U(\mathcal{U}[0, 1], \tilde{P}) = 1/S.$$

For an arbitrary P with distribution function G , the optimal approximation is given (by virtue of the quantile transform) by

$$z_s = G^{-1}\left(\frac{2s-1}{2S}\right), \quad s = 1, \dots, S,$$

$$p_s = 1/S, \quad s = 1, \dots, S.$$

The distance is bounded by $1/S$. If G is continuous, then the distance is exactly equal to $1/S$. In this case, the uniform distance approximation problem leads always to approximating distributions, which put the same mass $1/S$ at all mass points. This implies that the tails of P are not well represented by \tilde{P} . This drawback is not present when minimizing the Kantorovich distance.

15.3.1.5 Heavy Tail Control

Minimizing the Kantorovich distance minimizes the difference in expectation for all Lipschitz functions. However, it may not lead to a good approximation for higher moments or even the variance. If a good approximation of higher moments is also required, the Fortet–Mourier distance may be used.

Another possibility, which seems to be even more adapted to control tails, is to choose the Wasserstein r -distance instead of the Kantorovich distance for some appropriately chosen $r > 1$.

To approximate tails and higher moments are necessary for many real-world applications especially in the area of financial management.

Example Suppose we want to approximate the t -Student distribution P with 2 degrees of freedom, i.e., density $(2+x^2)^{-3/2}$ by a discrete distribution sitting on 5 points. The optimal approximation with the uniform distance is

$$\tilde{P}_1 = \left[\begin{array}{ccccc} 0.2 & 0.2 & 0.2 & 0.2 & 0.2 \\ -1.8856 & -0.6172 & 0 & 0.6172 & 1.8856 \end{array} \right],$$

while minimizing the Kantorovich distance, the approximated distribution is

$$\tilde{P}_2 = \left[\begin{array}{ccccc} 0.0446 & 0.2601 & 0.3906 & 0.2601 & 0.0446 \\ -4.58 & -1.56 & 0 & 1.56 & 4.58 \end{array} \right].$$

These results can be compared visually in Fig. 15.2 where one can see clearly that the minimization of the Kantorovich distance leads to a much better approximation of the tails.

But even this approximation can be improved by heavy tail control: Applying the stretching function χ_2 to the t -distribution, one gets the distribution with density (Fig. 15.3)

$$\begin{cases} (2+x^2)^{-3/2} & \text{if } |x| \leq 1, \\ \frac{1}{2\sqrt{x}}(2+|x|)^{-3/2} & \text{if } |x| > 1. \end{cases}$$

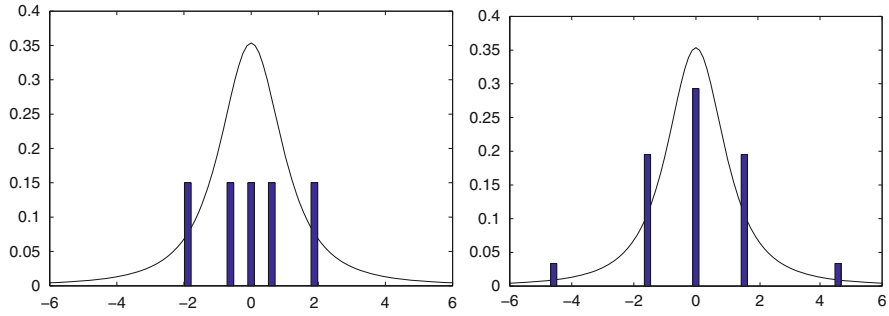


Fig. 15.2 Approximation of the $t(2)$ -distribution: uniform (left) and Kantorovich (right)

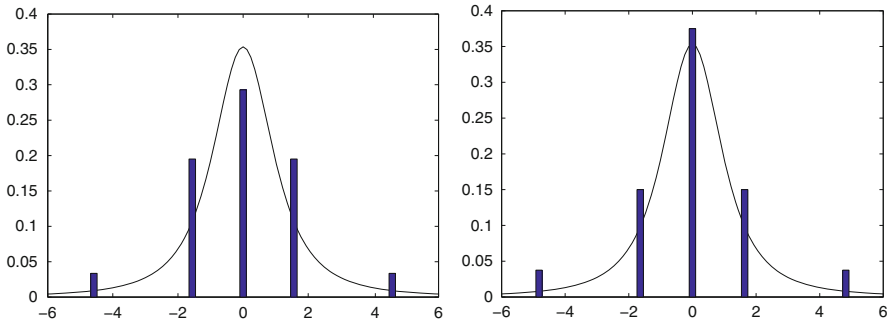


Fig. 15.3 Approximation of the $t(2)$ -distribution: Kantorovich with Euclidean distance (left) and Kantorovich with distorted distance d_{χ_2} (right)

This distribution has heavier tails than the original t -distribution. We approximate this distribution w.r.t the Kantorovich distance by a 5-point distribution and transform back the obtained mass points using the transformation $\chi_{1/2}$ to get

$$\tilde{P} = \left[\begin{array}{ccccc} 0.05 & 0.2 & 0.5 & 0.2 & 0.05 \\ -4.8216 & -1.6416 & 0 & 1.6416 & 4.8216 \end{array} \right].$$

Compared to the previous ones, this approximation has a better coverage of the tails. The second moment is better approximated as well.

15.3.2 Quasi-Monte Carlo

As above, let $\mathcal{U}[0, 1]^M$ be the uniform distribution on $[0, 1]^M$ and let $\tilde{P}_S = \frac{1}{S} \sum_{s=1}^S \delta_{z_s}$. Then the uniform distance $d_U(\mathcal{U}[0, 1]^M, \tilde{P}_S)$ is called the *star discrepancy* $D_S^*(z_1, \dots, z_S)$ of (z_1, \dots, z_S)

$$D_S^*(z_1, \dots, z_S) = \sup \left\{ \frac{1}{S} \sum_{s=1}^S \mathbb{1}_{\{[0, a_1] \times \dots \times [0, a_M]\}}(z_s) - \prod_{m=1}^M a_m : 0 < a_m \leq 1 \right\}. \tag{15.29}$$

A sequence $z = (z_1, z_2, \dots)$ in $[0, 1]^M$ is called a *low-discrepancy sequence*, if

$$\limsup_{S \rightarrow \infty} \frac{S}{\log(S)^M} D_S^*(z_1, \dots, z_S) < \infty,$$

i.e., if there is a constant C such that for all S ,

$$D_S^*(z) = D_S^*(z_1, \dots, z_S) \leq C \frac{\log(S)^M}{S}.$$

A low-discrepancy sequence allows the approximation of the uniform distribution in such a way that

$$d_U(\mathcal{U}[0, 1]^M, \tilde{P}_S) \leq C \frac{\log(S)^M}{S}.$$

Low-discrepancy sequences exist. The most prominent ones are as follows:

- The van der Corput sequence (for $M = 1$):
Fix a prime number p and let, for every s

$$s = \sum_{k=0}^L d_k(s) p^k$$

be the representation of n in the p -adic system. The sequence z_n is defined as

$$z_n^{(p)} = \sum_{k=0}^L d_k(s) p^{-k-1}.$$

- The Halton sequence (for $M > 1$):
Let p_1, \dots, p_M be different prime numbers. The sequence is

$$(z_s^{(1)}, \dots, z_s^{(M)}),$$

where $z_s^{(p_m)}$ are van der Corput sequences.

- The Hammersley sequence:
For a given Halton sequence $(z_s^{(1)}, \dots, z_s^{(M)})$, the Hammersley sequence enlarges it to \mathbb{R}^{M+1} by

$$(z_s^{(1)}, \dots, z_s^{(M)}, s/S).$$

- The Faure sequence.
- The Sobol sequence.
- The Niederreiter sequence.

(For a thorough treatment of QMC sequences, see Niederreiter (1992).) The quasi-Monte Carlo method was used in stochastic programming by Pennanen (2009). It is especially good if the objective is the expectation, but less advisable if tail or rare event properties of the distribution enter the objective or the constraints.

15.3.2.1 Comparing Different Quantizations

We¹ have implemented the following portfolio optimization problem: Maximize the expected average value at risk ($\mathbb{AV@R}$, CVaR) of the portfolio return under the constraint that the expected return exceeds some threshold and a budget constraint. We used historic data of 13 assets. The “true distribution” is their weekly performance for 10 consecutive years. The original data matrix of size 520×13 was reduced by scenario generation (scenario approximation) to a data matrix of size 50×13 . We calculated the optimal asset allocation for the original problem (1) and for the scenario methods: (2) minimal Kantorovich distance, (3) moment matching, and (4) quasi-Monte Carlo with a Sobol sequence. The Kantorovich distance was superior to all others, see Fig. 15.4.

15.3.3 Monte Carlo

The Monte Carlo approximation is a random approximation, i.e., \tilde{P} is a random measure and the distance $d(P, \tilde{P})$ is a random variable. This distinguishes the Monte Carlo method from the other methods.

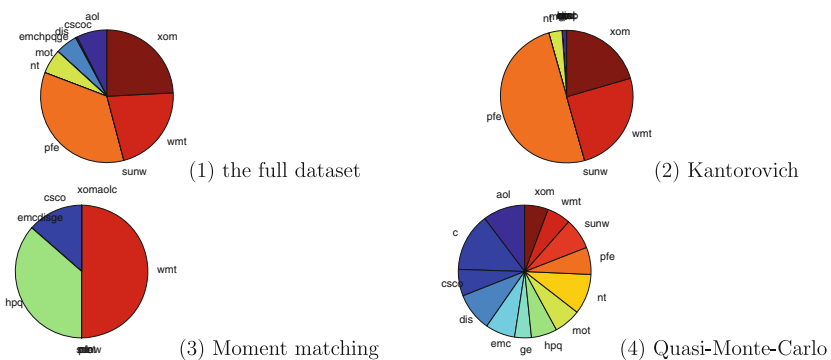


Fig. 15.4 Asset allocation using different scenario approximation methods. The pie charts show the optimal asset mix

¹ The programming of these examples was done by R. Hochreiter.

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

$$\tilde{P}_S = \frac{1}{S} \sum_{s=1}^S \delta_{X_s}.$$

15.3.3.1 The Uniform Distance of the MC Approximation

Let X_1, X_2, \dots be an i.i.d. sequence from a uniform $[0,1]$ distribution. Then by the famous Kolmogorov–Smirnov theorem

$$\lim_{S \rightarrow \infty} P\{\sqrt{S}D_S^*(X_1, \dots, X_S) > t\} = 2 \sum_{k=1}^{\infty} (-1)^{k-1} \exp(-2k^2 t^2).$$

15.3.3.2 The Kantorovich Distance of the MC Approximation

Let X_1, X_2, \dots be an i.i.d. sequence from a distribution with density g . Then

$$\lim_{S \rightarrow \infty} P\{S^{1/M} d_{KA}(P, \tilde{P}_S) \leq t\} = \int [1 - \exp(-t^M b_M g(x))] g(x) dx,$$

where $b_M = \frac{2\pi^{M/2}}{M\Gamma(M/2)}$ is the volume of the M -dimensional Euclidean ball (see Graf and Luschgy (2000), theorem 9.1).

15.4 Distances Between Multi-period Probability Measures

Let P be a distribution on

$$\mathbb{R}^{MT} = \underbrace{\mathbb{R}^M \times \mathbb{R}^M \times \dots \times \mathbb{R}^M}_{T \text{ times}}.$$

We interpret the parameter $t = 1, \dots, T$ as time and the parameter $m = 1, \dots, M$ as the multivariate dimension of the data.

While the multivariate dimension M does not add to much complexity, the multi-period dimension T is of fundamentally different nature. Time is the most important source of uncertainty and the time evolution gives increasing information. We consider the conditional distributions given the past and formulate scenario approximations in terms of the conditional distributions.

15.4.1 Distances Between Transition Probabilities

Any probability measure P on \mathbb{R}^{MT} can be dissected into the chain of transition probabilities

$$P = P_1 \circ P_2 \circ \dots \circ P_T,$$

where P_1 is the unconditional probability on \mathbb{R}^M for time 1 and P_t is the conditional probability for time t , given the past. We do not assume that P is Markovian, but bear in mind that for Markovian processes, the past reduces to the value one period before.

We will use the following notation: If $u = (u_1, \dots, u_T) \in \mathbb{R}^{MT}$, with $u_t \in \mathbb{R}^M$, then $u^t = (u_1, \dots, u_t)$ denotes the history of u up to time t .

The chain of transition probabilities are defined by the following relation: $P = P_1 \circ P_2 \circ \dots \circ P_T$, if and only if for all sequence of Borel sets A_1, A_2, \dots, A_T ,

$$P(A_1 \times \dots \times A_T) = \int_{A_1} \dots \int_{A_T} P_T(du_T | u^{T-1}) \dots P_3(du_3 | u^2) P_2(du_2 | u_1) P_1(du_1).$$

We assume that d makes \mathbb{R}^M a complete separable metric space, which implies that the existence of regular conditional probabilities is ensured (see, e.g., Durrett (1996), ch. 4, theorem 1.6): Recall that if R_1 and R_2 are metric spaces and \mathcal{F}_2 is a σ -algebra in R_2 , then a family of regular conditional probabilities is a mapping $(u, A) \mapsto P(A|u)$, where $u \in R_1, A \in \mathcal{F}_2$, which satisfies

- $u \mapsto P(A|u)$ is (Borel) measurable for all $A \in \mathcal{F}_2$, and
- $A \mapsto P(A|u)$ is a probability measure on \mathcal{F}_2 for each $u \in R_1$.

We call regular conditional probabilities $P(A|u)$ also *transition probabilities*.

The notion of probability distances may be extended to transition probabilities: Let P and Q be two transition probabilities on $R_1 \times \mathcal{F}_2$ and let d be some probability distance. Define

$$d(P, Q) := \sup_u d(P(\cdot|u), Q(\cdot|u)). \tag{15.30}$$

In this way, all probability distances like d_U, d_{KA}, d_{FM} may be extended to transition probabilities in a natural way.

The next result compares the Wasserstein distance $d_r(P, Q)$ with the distances of the components $d_r(P_t, Q_t)$. To this end, introduce the following notation. Let d be some distance on \mathbb{R}^M . For a vector $w = (w_1, \dots, w_T)$ of nonnegative weights we introduce the distance $d^t(u^t, v^t)^r := \sum_{s=1}^t w_s \cdot d(u_s, v_s)^r$ on \mathbb{R}^{tM} for all t .

Then the following theorem holds true:

Theorem 3.1 Assume that P and Q are probability measures on \mathbb{R}^{TM} , such that for all conditional measures P_t resp. Q_t we have that

$$d_r(P_{t+1}[\cdot|u^t], Q_{t+1}[\cdot|v^t]) \leq \tau_{t+1} + K_{t+1} \cdot d^t(u^t, v^t) \tag{15.31}$$

for $t = 1, \dots, T - 1$. Then

$$\mathbf{d}_r(P^{t+1}, Q^{t+1})^r \leq w_{t+1}\tau_{t+1} + (1 + w_{t+1}K_{t+1})\mathbf{d}_r(P^t, Q^t)^r \quad (15.32)$$

and therefore

$$\mathbf{d}_r(P, Q)^r \leq \mathbf{d}_r(P_1, Q_1)^r \prod_{s=1}^T (1 + w_s K_s) + \sum_{s=1}^T \tau_s w_s \prod_{j=1}^s (1 + w_j K_j). \quad (15.33)$$

Proof Let π^t denote the optimal joint measure (optimal transportation plan) for $\mathbf{d}_r(P^t, Q^t)$ and let $\pi_{t+1}[\cdot|u, v]$ denote the optimal joint measure for $\mathbf{d}_r(P_{t+1}[\cdot|u], Q_{t+1}[\cdot|v])$ and define the combined measure

$$\pi^{t+1}[A_t \times B^t, C_t \times D^t] := \int_{A_t \times C_t} \pi_{t+1}[B^t \times D^t | u^t, v^t] \pi^t[du^t, dv^t].$$

Then

$$\begin{aligned} \mathbf{d}_r(P^{t+1}, Q^{t+1})^r &\leq \int d^{t+1}(u^{t+1}, v^{t+1})^r \pi^{t+1}[du^{t+1}, dv^{t+1}] \\ &= \int [d^t(u^t, v^t)^r + w_{t+1}d_{t+1}(u_{t+1}, v_{t+1})^r] \\ &\quad \pi_t[du_{t+1}, dv_{t+1}|u^t, v^t] \pi^t[du^t, dv^t] \\ &= \int d^t(u^t, v^t)^r \pi^t[du^t, dv^t] + \\ &\quad + w_{t+1} \int d_{t+1}(u_{t+1}, v_{t+1})^r \pi_t[du_{t+1}, dv_{t+1}|u^t, v^t] \pi^t[du^t, dv^t] \\ &= \int d^t(u^t, v^t)^r \pi^t[du^t, dv^t] + \\ &\quad + w_{t+1} \int \mathbf{d}_r(P_{t+1}[\cdot|u^t], Q_{t+1}[\cdot|v^t])^r \pi^t[du^t, dv^t] \\ &\leq \int d^t(u^t, v^t)^r \pi^t[du^t, dv^t] + \\ &\quad + w_{t+1} \int \tau_{t+1} + K_{t+1}d^t(u^t, v^t)^r \pi^t[du^t, dv^t] \\ &= \mathbf{d}_r(P^t, Q^t)^r + w_{t+1}(\tau_{t+1} + K_{t+1})\mathbf{d}_r(P^t, Q^t)^r \\ &= w_{t+1}\tau_{t+1} + (1 + w_{t+1}K_{t+1})\mathbf{d}_r(P^t, Q^t)^r. \end{aligned}$$

This demonstrates (15.32). Applying this recursion for all t leads to (15.33). \square

Definition 3.2 Let P be a probability on \mathbb{R}^{MT} , dissected into transition probabilities P_1, \dots, P_T . We say that P has the (K_2, \dots, K_T) -Lipschitz property, if for $t = 2, \dots, T$,

$$\mathbf{d}_r(P_t(\cdot|u), P_t(\cdot|v)) \leq K_t d(u, v).$$

Example (The normal distribution has the Lipschitz property) Consider a normal distribution P on $\mathbb{R}^{m_1+m_2}$, i.e.,

$$P = \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix}\right).$$

The conditional distribution, given $u \in \mathbb{R}^{m_1}$, is

$$P_2(\cdot|u) = \mathcal{N}\left(\mu_1 - \Sigma_{12}\Sigma_{11}^{-1}(u - \mu_1), \Sigma_{22} - \Sigma_{12}\Sigma_{11}\Sigma_{12}^{-1}\right).$$

We claim that

$$d_{KA}(P_2(\cdot|u), P_2(\cdot|v)) \leq \|\Sigma_{12}\Sigma_{11}^{-1}(u - v)\|. \tag{15.34}$$

Let $Y \sim \mathcal{N}(0, \Sigma)$ and let $Y_1 = a_1 + Y$ and $Y_2 = a_2 + Y$. Then $Y_1 \sim \mathcal{N}(a_1, \Sigma)$, $Y_2 \sim \mathcal{N}(a_2, \Sigma)$ and therefore

$$d_{KA}(\mathcal{N}(a_1, \Sigma), \mathcal{N}(a_2, \Sigma)) \leq \mathbb{E}(\|Y_1 - Y_2\|) = \|a_1 - a_2\|.$$

Setting $a_1 = \mu_1 - \Sigma_{12}\Sigma_{11}^{-1}(u - \mu_1)$, $a_2 = \mu_1 - \Sigma_{12}\Sigma_{11}^{-1}(v - \mu_1)$ and $\Sigma = \Sigma_{22} - \Sigma_{12}\Sigma_{11}\Sigma_{12}^{-1}$, the result (15.34) follows.

Corollary 3.3 Condition (15.31) is fulfilled, if the following two conditions are met:

- (i) P has the (K_2, \dots, K_T) -Lipschitz property in the sense of Definition 3.2
- (ii) Closeness of all P_t and Q_t in the following sense:

$$d_r(P_t, Q_t) = \sup_u d_r(P_t(\cdot|u), Q_t(\cdot|u)) \leq \tau_t$$

The Lipschitz assumption for P is crucial. Without this condition, no relation between $d_r(P, Q)$ and the $d_r(P_t, Q_t)$'s holds. In particular, the following example shows that $d_1(P_t, Q_t)$ may be arbitrarily small and yet $d_1(Q, P)$ is far away from zero.

Example Let $P = P_1 \circ P_2$ be the measure on \mathbb{R}^2 , such that

$$P_1 = \mathcal{U}[0, 1],$$

$$P_2(\cdot|u) = \begin{cases} \mathcal{U}[0, 1] & \text{if } u \in \mathbb{Q}, \\ \mathcal{U}[1, 2] & \text{if } u \in \mathbb{R} \setminus \mathbb{Q}, \end{cases}$$

where $\mathcal{U}[a, b]$ denotes the uniform distribution on $[a, b]$. Obviously, P is the uniform distribution on $[0, 1] \times [1, 2]$. (P_1, P_2) has no Lipschitz property. Now, let $Q^{(n)} = Q_1^{(n)} \circ Q_2^{(n)}$, where

$$Q_1^{(n)} = \sum_{k=1}^n \frac{1}{n} \delta_{k/n},$$

$$Q_2^{(n)}(\cdot|u) = \begin{cases} \sum_{k=1}^n \frac{1}{n} \delta_{k/n} & \text{if } u \in \mathbb{Q}, \\ \sum_{k=1}^n \frac{1}{n} \delta_{1+k/n} & \text{if } u \in \mathbb{R} \setminus \mathbb{Q}. \end{cases}$$

Note that $Q^{(n)} = Q_1^{(n)} \circ Q_2^{(n)}$ converges weakly to the uniform distribution on $[0, 1] \times [0, 1]$. In particular, it does not converge to P and the distance is $d_1(P, Q^{(n)}) = 1$. However,

$$d_1(P_1, Q_1^{(n)}) = 1/n; \quad d_1(P_2, Q_2^{(n)}) = 1/n.$$

15.4.2 Filtrations and Tree Processes

In many stochastic decision problems, a distinction has to be made between the scenario process (which contains decision-relevant values, such as prices, demands, and supplies) and the information which is available. Here is a simple example due to Philippe Artzner.

Example A fair coin is tossed three times. Situation A: The payoff process $\xi^{(A)}$ is

$$\xi_1^{(A)} = 0; \quad \xi_2^{(A)} = 0;$$

$$\xi_3^{(A)} = \begin{cases} 1 & \text{if heads is shown at least two times,} \\ 0 & \text{otherwise.} \end{cases}$$

We compare this process to another payoff process (situation B)

$$\xi_1^{(B)} = 0, \quad \xi_2^{(B)} = 0,$$

$$\xi_3^{(B)} = \begin{cases} 1 & \text{if heads is shown at least the last throw,} \\ 0 & \text{otherwise.} \end{cases}$$

The available information is relevant: It makes a difference, whether we may observe the coin tossing process or not. Suppose that the process is observable. Then, after the second throw, we know the final payoff, if two times the same side appeared. Thus our information is complete although the experiment has not ended yet. Only in case that two different sides appeared, we have to wait until the last throw to get the final information.

Mathematically, the information process is described by a filtration.

Definition 3.4 Let (Ω, \mathcal{F}, P) be probability space. A *filtration* \mathcal{F} on this probability space is an increasing sequence of σ -fields $\mathcal{F} = (\mathcal{F}_1, \dots, \mathcal{F}_T)$ where for all t ,

$\mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F}$. We may always add the trivial σ -field $\mathcal{F}_0 = \{\emptyset, \Omega\}$ as the zeroth element of a filtration.

A scenario process $\xi = (\xi_1, \dots, \xi_T)$ is *adapted* to the filtration \mathcal{F} , if ξ_t is \mathcal{F}_t measurable for $t = 1, \dots, T$. If ξ_t is measurable w.r.t. \mathcal{F}_t , we use the notation

$$\xi_t \triangleleft \mathcal{F}_t;$$

if the process ξ is adapted to \mathcal{F} , we use the similar notation

$$\xi \triangleleft \mathcal{F}.$$

The general situation is as follows: We separate the information process from the value process. Information is modeled by a filtration \mathcal{F} , while the scenario values are modeled by a (real or vector values process) ξ , which is adapted to \mathcal{F} . We call a pair (ξ, \mathcal{F}) a *process-and-information pair*. In general, the filtration \mathcal{F} is larger than the one generated by the process ξ .

In finite probability spaces, there is a one-to-one correspondence between filtrations and probability trees. A process is adapted to such a filtration, iff it is a function of the nodes of the tree. Figures 15.5 and 15.6 show the tree processes and the adapted scenario process sitting on the tree for Artzner's example and situations (A) and (B).

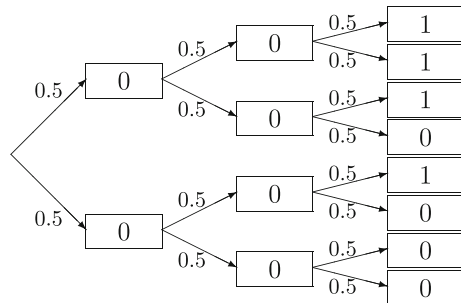


Fig. 15.5 The payoff process $\xi^{(A)}$ sitting on the tree (the filtration) generated by the coin process

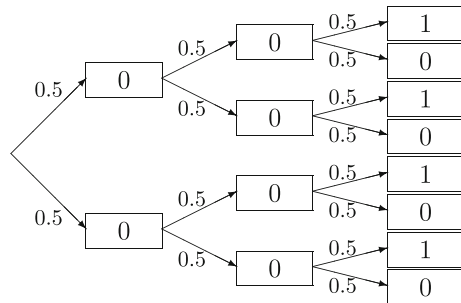


Fig. 15.6 The payoff process $\xi^{(B)}$ sitting on the tree (the filtration) generated by the coin process

15.4.2.1 Tree Processes

Definition 3.5 A stochastic process $\nu = (\nu_1, \dots, \nu_T)$ is called a *tree process*, if $\sigma(\nu_1), \sigma(\nu_2), \dots, \sigma(\nu_T)$ is a filtration. A tree process is characterized by the property that the conditional distribution of ν_1, \dots, ν_t given ν_{t+1} is degenerated, i.e., concentrated on a point. Notice that any stochastic process $\eta = (\eta_1, \dots, \eta_T)$ can be transformed into a tree process by considering its history process

$$\begin{aligned} \nu_1 &= \eta_1, \\ \nu_2 &= (\eta_1, \eta_2), \\ &\vdots \\ \nu_T &= (\eta_1, \dots, \eta_T). \end{aligned}$$

Let (ν_t) be a tree process and let $\mathcal{F} = \sigma(\nu)$. It is well known that any process ξ adapted to \mathcal{F} can be written as $\xi_t = f_t(\nu_t)$ for some measurable function f_t . Thus a stochastic program may be formulated using the tree process ν and the functions f_t .

Let ν be a tree process according to Definition 3.5. We assume throughout that the state spaces of all tree processes are Polish. In this case, the joint distribution P of ν_1, \dots, ν_T can be factorized into the chain of conditional distributions, which are given as transition kernels, that is

$$\begin{aligned} P_1(A) &= P\{\nu_1 \in A\}, \\ P_2(A|u) &= P\{\nu_2 \in A | \nu_1 = u\}, \\ &\vdots \\ P_t(A|u_1, \dots, u_{t-1}) &= P\{\nu_t \in A | \nu_1 = u_1, \dots, \nu_{t-1} = u_{t-1}\}. \end{aligned}$$

Definition 3.6 (Equivalence of tree processes) Let ν resp. $\bar{\nu}$ two tree processes, which are defined on possibly different probability spaces (Ω, \mathcal{F}, P) resp. $(\bar{\Omega}, \bar{\mathcal{F}}, \bar{P})$. Let P_1, P_2, \dots, P_T resp. $\bar{P}_1, \bar{P}_2, \dots, \bar{P}_T$ be the pertaining chain of conditional distributions. The two tree processes are *equivalent*, if the following properties hold: There are bijective functions k_1, \dots, k_T mapping the state spaces of ν to the state spaces of $\bar{\nu}$, such that the two processes ν_1, \dots, ν_T and $k_1^{-1}(\bar{\nu}_1), \dots, k_T^{-1}(\bar{\nu}_T)$ coincide in distribution.

Notice that for tree processes defined on the same probability space, one could define equivalence simply as $\nu_t = k_t(\bar{\nu}_t)$ a.s. The value of Definition 3.6 is that it extends to tree processes on different probability spaces.

Definition 3.7 (Equivalence of scenario process-and-information pairs) Let (ξ, ν) be a scenario process-and-information pair, consisting of a scenario process ξ and a tree process, such that $\xi \triangleleft \sigma(\nu)$. Since $\xi \triangleleft \sigma(\nu)$, there are measurable functions f_t such that

$$\xi_t = f_t(\nu_t), \quad t = 1, \dots, T.$$

Two pairs (ξ, ν) and $(\bar{\xi}, \bar{\nu})$ are called *equivalent*, if ν and $\bar{\nu}$ are equivalent in the sense of Definition 3.6 with equivalence mappings k_t such that

$$\bar{\xi}_t = f_t(k_t^{-1}(\xi_t)) \quad \text{a.s.}$$

We agree that equivalent process-and-information pairs will be treated as equal. Notice that by drawing a finite tree with some values sitting on its nodes we automatically mean an equivalence class of process-and-information pairs.

If two stochastic optimization problems are defined in the same way, but using equivalent process-and-information pairs, then they have equivalent solutions. It is easy to see that an optimal solution of the first problem can be transformed into an optimal solution of the second problem via the equivalence functions k_t .

Figures 15.7 and 15.8 illustrate the concept of equivalence. The reader is invited to find the equivalence functions k_t in this example.

15.4.2.2 Distances Between Filtrations

In order to measure the distances between two discretizations on the same probability space one could first measure the distance of the respective filtrations. This

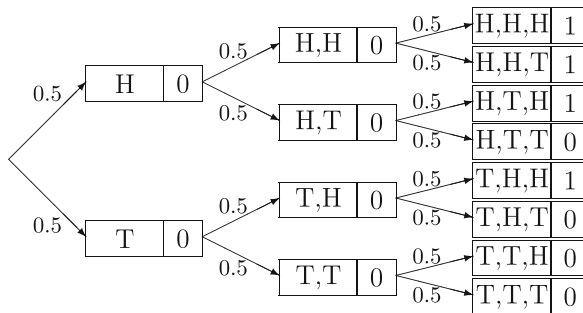


Fig. 15.7 Artzner's example: The tree process is coded by the history of the coin throws (H = head, T = tail), the payoffs are functions of the tree process

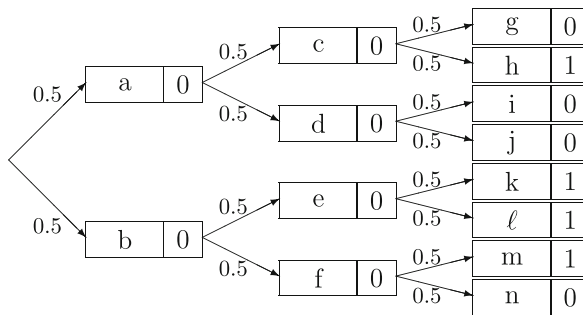


Fig. 15.8 A process-and-information pair which is equivalent to Artzner's example in Fig. 15.4. The coding of the tree process is just by subsequent letters

concept of filtration distance for scenario generation has been developed by Heitsch and Roemisch (2003) and Heitsch et al. (2006).

Let $(\Omega, \tilde{\mathcal{F}}, P)$ be a probability space and let \mathcal{F} and $\tilde{\mathcal{F}}$ be two sub-sigma algebras of $\tilde{\mathcal{F}}$. Then the *Kudo* distance between \mathcal{F} and $\tilde{\mathcal{F}}$ is defined as

$$\mathfrak{D}(\mathcal{F}, \tilde{\mathcal{F}}) = \max(\sup\{\|\mathbb{1}_A - \mathbb{E}(\mathbb{1}_A|\tilde{\mathcal{F}})\|_p : A \in \mathcal{F}\}, \sup\{\|\mathbb{1}_B - \mathbb{E}(\mathbb{1}_B|\mathcal{F})\|_p : A \in \tilde{\mathcal{F}}\}). \tag{15.35}$$

The most important case is the distance between a sigma algebra and a sub-sigma algebra of it: If $\tilde{\mathcal{F}} \subseteq \mathcal{F}$, then (15.35) reduces to

$$\mathfrak{D}(\mathcal{F}, \tilde{\mathcal{F}}) = \sup\{\|\mathbb{1}_A - \mathbb{E}(\mathbb{1}_A|\tilde{\mathcal{F}})\|_p : A \in \mathcal{F}\}. \tag{15.36}$$

If $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \dots)$ is an infinite filtration, one may ask, whether this filtration converges to a limit, i.e., whether there is a \mathcal{F}_∞ , such that

$$\mathfrak{D}(\mathcal{F}_t, \mathcal{F}_\infty) \rightarrow 0 \quad \text{for } t \rightarrow \infty.$$

If \mathcal{F}_∞ is the smallest σ -field, which contains all \mathcal{F}_t , then $\mathfrak{D}(\mathcal{F}_t, \mathcal{F}_\infty) \rightarrow 0$. However, the following example shows that not every discrete increasing filtration converges to the Borel σ -field.

Example Let \mathcal{F} be the Borel-sigma algebra on $[0,1)$ and let $\tilde{\mathcal{F}}_n$ be the sigma algebra generated by the sets $\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)$, $k = 0, \dots, 2^n - 1$. Moreover, let $A_n = \bigcup_{k=0}^{2^n-1} \left[\frac{k}{2^n}, \frac{2k+1}{2^{n+1}}\right)$. Then $\mathbb{E}(\mathbb{1}_{A_n}|\tilde{\mathcal{F}}_n) = \frac{1}{2}$ and $\|\mathbb{1}_{A_n} - \mathbb{E}(\mathbb{1}_{A_n}|\tilde{\mathcal{F}}_n)\|_p = 1/2$ for all n . While one has the intuitive feeling that $\tilde{\mathcal{F}}_n$ approaches \mathcal{F} , the distance is always $1/2$.

15.4.3 Nested Distributions

We define a nested distribution as the distribution of a process-and-information pair, which is invariant w.r.t. to equivalence. In the finite case, the nested distribution is the distribution of the scenario tree or better of the class of equivalent scenario trees.

Let d be a metric on \mathbb{R}^M , which makes it a complete separable metric space and let $\mathcal{P}_1(\mathbb{R}^M, d)$ be the family of all Borel probability measures P on (\mathbb{R}^M, d) such that

$$\int d(u, u_0) dP(u) < \infty \tag{15.37}$$

for some $u_0 \in \mathbb{R}^M$.

For defining the nested distribution of a scenario process (ξ_1, \dots, ξ_T) with values in \mathbb{R}^M , we define the following spaces in a recursive way:

$$\mathfrak{E}_1 := \mathbb{R}^M,$$

with distance $d_1(u, v) = d(u, v)$:

$$\mathfrak{E}_2 := \mathbb{R}^M \times \mathcal{P}_1(\mathfrak{E}_1)$$

with distance $d_2((u, P), (u, Q)) := d(u, v) + d_{KA}(P, Q; d_1)$:

$$\mathfrak{E}_3 := \mathbb{R}^M \times \mathcal{P}_1(\mathfrak{E}_2) = (\mathbb{R}^M \times \mathcal{P}_1(\mathbb{R}^M \times \mathcal{P}_1(\mathbb{R}^M)))$$

with distance $d_3((u, \mathbb{P}), (v, \mathbb{Q})) := d(u, v) + d_{KA}(\mathbb{P}, \mathbb{Q}; d_2)$. This construction is iterated until

$$\mathfrak{E}_T = \mathbb{R}^M \times \mathcal{P}_1(\mathfrak{E}_{T-1})$$

with distance $d_T((u, \mathbb{P}), (v, \mathbb{Q})) := d(u, v) + d_{KA}(\mathbb{P}, \mathbb{Q}; d_{T-1})$.

Definition 3.8 (Nested distribution) A Borel probability distribution \mathbb{P} with finite first moments on \mathfrak{E}_T is called a *nested distribution of depth T*; the distance d_T is called the *nested distance* (Pflug 2009).

In discrete cases, a nested distribution may be represented as a recursive structure: Recall that we represent discrete probabilities by a list of probabilities (in the first row) and the values (in the subsequent rows).

Notice that *one* value or vector of values may only carry *one* probability, the following structure on the left does not represent a valid distribution, the structure right is correct:

$$\left[\begin{array}{cccc} 0.1 & 0.2 & 0.4 & 0.3 \\ \hline 3.0 & 3.0 & 1.0 & 5.0 \end{array} \right] \quad \left[\begin{array}{ccc} 0.3 & 0.4 & 0.3 \\ \hline 3.0 & 1.0 & 5.0 \end{array} \right].$$

In the same manner, but in a recursive way, we may now represent a nested distribution as a structure, where some values are distributions themselves:

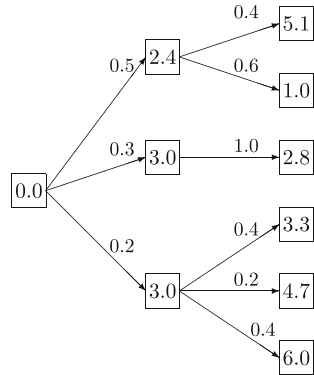
$$\left[\begin{array}{ccc} 0.2 & 0.3 & 0.5 \\ \hline 3.0 & 3.0 & 2.4 \\ \left[\begin{array}{ccc} 0.4 & 0.2 & 0.4 \\ \hline 6.0 & 4.7 & 3.3 \end{array} \right] & \left[\begin{array}{c} 1.0 \\ \hline 2.8 \end{array} \right] & \left[\begin{array}{cc} 0.6 & 0.4 \\ \hline 1.0 & 5.1 \end{array} \right] \end{array}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

This recursive structure encodes the tree shown in Fig. 15.9.

For any nested distribution \mathbb{P} , there is an embedded multivariate distribution P . We explain the projection from the nested distribution to the embedded multivariate distribution just for the depth 3, the higher depths being analogous:

Let \mathbb{P} be a nested distribution on \mathfrak{E}_3 , which has components η_1 (a real random variable) and μ_2 (a random distribution on \mathfrak{E}_2 , which means that it has in turn components η_2 (a random variable) and μ_3 , a random distribution on \mathbb{R}^m).

Fig. 15.9 Example tree



Then the pertaining multivariate distribution on $\mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R}^m$ is given by

$$P(A_1 \times A_2 \times A_3) = \mathbb{E}_{\mathbb{P}}[\mathbf{1}_{\{\eta_1 \in A_1\}} \mathbb{E}_{\mu_2}(\mathbf{1}_{\{\eta_2 \in A_2\}} \mu_2(A_3))]$$

for A_1, A_2, A_3 Borel sets in \mathbb{R}^m .

The mapping from the nested distribution to the embedded distribution is not injective: There may be many different nested distributions which all have the same embedded distribution, see Artzner's example: $\xi^{(A)}$ and $\xi^{(B)}$ in Figs. 15.5 and 15.6 have different nested distributions, but the same embedded distributions. Here is another example:

Example Consider again the process-and-information pair (the tree) of Fig. 15.9. The embedded multivariate but non-nested distribution of the scenario process is

$$\begin{bmatrix} 0.08 & 0.04 & 0.08 & 0.3 & 0.3 & 0.2 \\ 3.0 & 3.0 & 3.0 & 3.0 & 2.4 & 2.4 \\ 6.0 & 4.7 & 3.3 & 2.8 & 1.0 & 5.1 \end{bmatrix}.$$

Evidently, this 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

$$\begin{bmatrix} \begin{matrix} 0.5 & 0.5 \\ 3.0 & 2.4 \end{matrix} \\ \begin{bmatrix} 0.16 & 0.08 & 0.16 & 0.6 \\ 6.0 & 4.7 & 3.3 & 2.8 \end{bmatrix} \begin{bmatrix} 0.6 & 0.4 \\ 1.0 & 5.1 \end{bmatrix} \end{bmatrix}$$

since in this case the two nodes with value 3.0 have to be identified.

Lemma 3.9 Let (ξ, ν) be equivalent to $(\bar{\xi}, \bar{\nu})$ in the sense of Definition 3.7. Then both pairs generate the same nested distribution.

For a proof see Pflug (2009).

15.4.3.1 Computing the Nested Distance

Recall that a nested distribution is the distribution of a process-and-information pair, which is invariant w.r.t. equivalence. In the finite case, the nested distribution is the distribution of the scenario tree (or better of the class of equivalent scenario trees), which was defined in a recursive way.

To compute the nested distance of two discrete nested distributions, P and \tilde{P} say, corresponds to computing the distance of two tree processes ν and $\tilde{\nu}$. We further elaborate the algorithmic approach here, which is inherent is the recursive definition of nested distributions and the nested distance.

To this end recall first that the linear program (15.20) was defined so general, allowing to compute the distance of general objects whenever probabilities p_s, p_t , and the cost matrix $c_{s,t}$ were specified – no further specifications of the samples z are needed. We shall exploit this fact; in the situation we describe here the samples z_s and z_t represent nested distributions with finitely many states, or equivalently entire trees.

Consider the nodes in the corresponding tree: Any node is either a terminal node, or a node with children, or the root node of the respective tree. The following recursive treatment is in line with this structure:

- **Recursion start** (discrete nested distribution of depth 1) In this situation the nested distributions themselves are elements of the underlying space, $P \in \Xi_T$ ($\tilde{P} \in \Xi_T$ resp.), for which we have that

$$d_T(P, \tilde{P}) = d(P, \tilde{P})$$

according to the definition.

In the notion of trees this situation corresponds to trees which consist of a single node only, which are terminal nodes (leaf nodes).

- **Recursive step** (discrete nested distribution of depth $t > 1$) In this situation the nested distribution may be dissected in

$$P_t = \left(n_t, \sum_s p_s^t \delta_{z_s^{t+1}} \right) \text{ and } \tilde{P}_t = \left(\tilde{n}_t, \sum_{\tilde{s}} \tilde{p}_{\tilde{s}}^t \delta_{z_{\tilde{s}}^{t+1}} \right),$$

where $z_s^{t+1} \in \Xi_{t+1}$ and $z_{\tilde{s}}^{t+1} \in \Xi_{t+1}$.

We shall assume that d_{t+1} is recursively available, thus define the matrix

$$c_{s,\tilde{s}} := d(n_t, \tilde{n}_t) + d_{t+1} \left(z_s^{t+1}, z_{\tilde{s}}^{t+1} \right)$$

and apply the linear program (15.20) together with the probabilities $p = (p_s)_s$ and $\tilde{p} = (\tilde{p}_{\tilde{s}})_{\tilde{s}}$, resulting with the distance $d_t(P_t, \tilde{P}_t)$.

In the notion of trees, again, n_t and \tilde{n}_t represent nodes with children. The children are given by the first components of z_s^{t+1} ($z_{\tilde{s}}^{t+1}$, resp.); further notice that

$(z_s^{t+1})_s$ are just all subtrees of n_t , all this information is encoded in the nested distribution P_t .

- **Recursive end** (nested distribution of depth $t = T$) The computation of the distance has already been established in the recursion step.

Note, that the corresponding trees in this situation represent the entire tree process here.

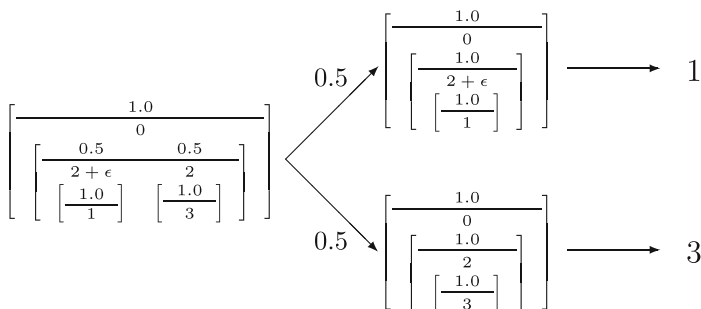
15.4.3.2 The Canonical Construction

Suppose that \mathbb{P} and $\tilde{\mathbb{P}}$ are nested distributions of depth T . One may construct the pertaining tree processes in a canonical way: The canonical name of each node is the nested distribution pertaining of the subtree, for which this node is the root. Two valued trees are equivalent, iff the respective canonical constructions are identical.

Example 3.10 Consider the following nested distribution:

$$\mathbb{P}_\epsilon = \left[\begin{array}{c} \frac{0.5 \quad 0.5}{2 \quad 2 + \mathbb{E}} \\ \left[\begin{array}{c} \frac{1.0}{3} \\ \frac{1.0}{1} \end{array} \right] \end{array} \right].$$

We construct a tree process such that the names of the nodes are the corresponding nested distributions of the subtrees:



The nested distance generates a much finer topology than the distance of the multivariate distributions.

Example (see Heitsch et al. (2006)) Consider the nested distributions of Example 3.10 and

$$\mathbb{P}_0 = \left[\begin{array}{c} \frac{1.0}{2} \\ \left[\begin{array}{c} 0.5 \quad 0.5 \\ 3 \quad 1 \end{array} \right] \end{array} \right].$$

Note that the pertaining multivariate distribution of \mathbb{P}_ϵ on \mathbb{R}^2 converges weakly to the one of \mathbb{P}_0 , if $\epsilon \rightarrow 0$. However, the nested distributions do not converge: The nested distance is $\mathbf{d}(\mathbb{P}_\epsilon, \mathbb{P}_0) = 1 + \epsilon$ for all ϵ .

Lemma 3.11 Let \mathbb{P} be a nested distribution, P its multivariate distribution, which is dissected into the chain $P = P_1 \circ P_2 \circ \dots \circ P_T$ of conditional distributions. If P has the (K_2, \dots, K_T) -Lipschitz property (see Definition 3.2), then

$$\mathbf{d}(\mathbb{P}, \tilde{\mathbb{P}}) \leq \sum_{t=1}^T \mathbf{d}_1(P_t, \tilde{P}_t) \prod_{s=t+1}^{T+1} (K_s + 1).$$

For a proof see Pflug (2009).

Theorem 3.12 Let \mathbb{P} resp. $\tilde{\mathbb{P}}$ be two nested distributions and let P resp. \tilde{P} be the pertaining multi-period distributions. Then

$$\mathbf{d}_1(P, \tilde{P}) \leq \mathbf{d}(\mathbb{P}, \tilde{\mathbb{P}}).$$

Thus the mapping, which maps the nested distribution to the embedded multi-period distribution, is Lipschitz(1). The topologies generated by the two distances are not the same: The two trees in Figs. 15.5 resp. 15.6 are at multivariate Kantorovich distance 0, but their nested distance is $\mathbf{d} = 0.25$.

Proof To find the Kantorovich distance between P and \tilde{P} one has to find a joint distribution for two random vectors ξ and $\tilde{\xi}$ such that the marginals are P resp. \tilde{P} . Among all these joint distributions, the minimum of $d(\xi_1, \tilde{\xi}_1) + \dots + d(\xi_T, \tilde{\xi}_T)$ equals $d(P, \tilde{P})$. The nested distance appears in a similar manner; the only difference is that for the joint distribution also the conditional marginals should be equal to those of \mathbb{P} resp. $\tilde{\mathbb{P}}$. Therefore, the admissible joint distributions for the nested distance are subsets of the admissible joint distributions for the distance between P and \tilde{P} and this implies the assertion. \square

Remark The advantage of the nested distance is that topologically different scenario trees with completely different values can be compared and their distance can be calculated. Consider the example of Fig. 15.10. The conditional probabilities $P(\cdot|u)$ and $Q(\cdot|v)$ are incomparable, since no values of u coincide with values of v . Moreover, the Lipschitz condition of Definition 3.2 does not make sense for trees and therefore Theorem 3.1 is not applicable. However, the nested distance is well defined for two trees.

By calculation, we find the nested distance $\mathbf{d} = 1.217$.

Example Here is an example of the nested distance between a continuous process and a scenario tree process. Let \mathbb{P} be the following continuous nested distribution: $\xi_1 \sim N(0, 1)$ and $\xi_2|\xi_1 \sim N(\xi_1, 1)$. Let $\tilde{\mathbb{P}}$ be the discrete nested distribution

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

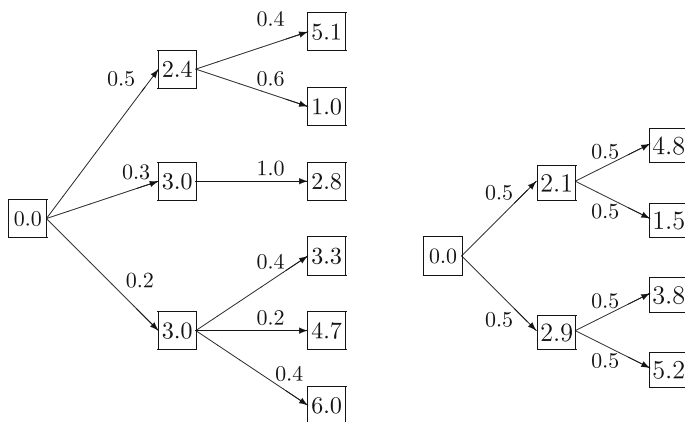


Fig. 15.10 Two valuated trees

Then the nested distance is $d(\mathbb{P}, \tilde{\mathbb{P}}) = 0.8215$, which we found by numerical distance calculation.

15.5 Scenario Tree Construction and Reduction

Suppose that a scenario process (ξ_t) has been estimated. In this section, we describe how to construct a scenario tree out of it. We begin with considering the case, where the information is the one which is generated by the scenario process and the method is based on the conditional probabilities. Later, we present the case where a tree process representing the information and a scenario process sitting on the tree process has been estimated.

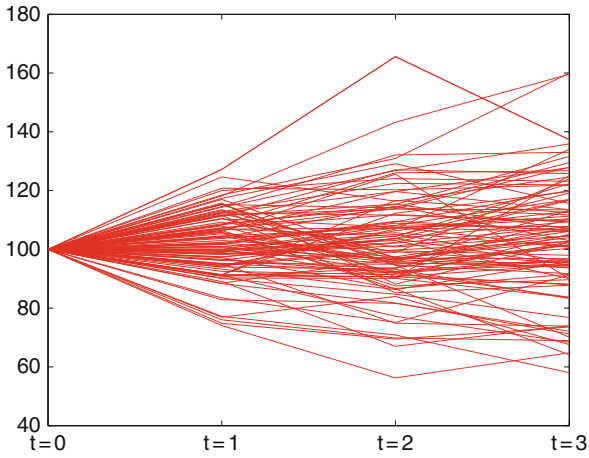
15.5.1 Methods Based on the Conditional Distance

The most widespread method to construct scenario trees is to construct them recursively from root to leaves. The discretization for the first stage is done as described in Section 15.2. Suppose a chosen value for the first stage is x . Then the reference process ξ_t is conditioned to the set $\{\omega : |\xi_1 - x| \leq \epsilon\}$, where ϵ is chosen in such a way that enough trajectories are contained in this set. The successor nodes of x are then chosen as if x were the root and the conditional process were the total process.

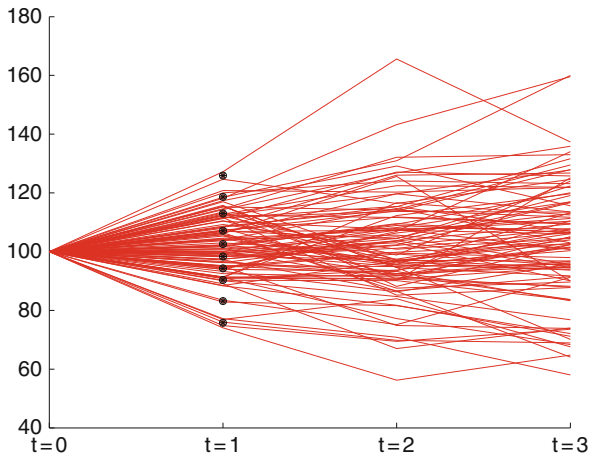
The original distributions may either be continuous (e.g., estimated distributions) or more commonly discrete (e.g., simulated with some random generation mechanism which provides some advantages for practical usage). Different solution techniques for continuous as well as discrete approximation have been outlined in Pflug (2001) and Hochreiter and Pflug (2007).

Example The typical algorithm for finding scenario trees based on the Kantorovich distance can be illustrated by the following pictures.

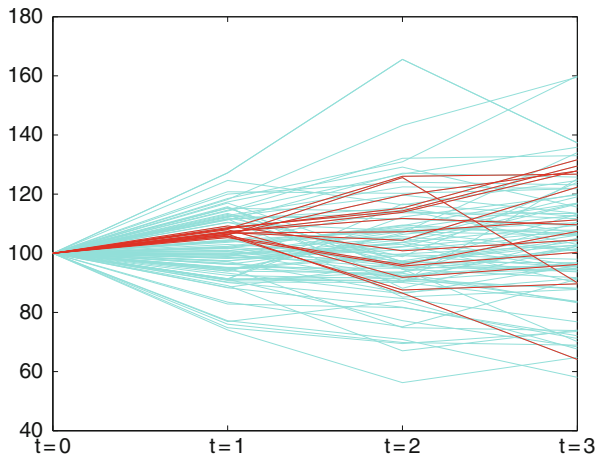
Step 1. Estimate the probability model and simulate future trajectories



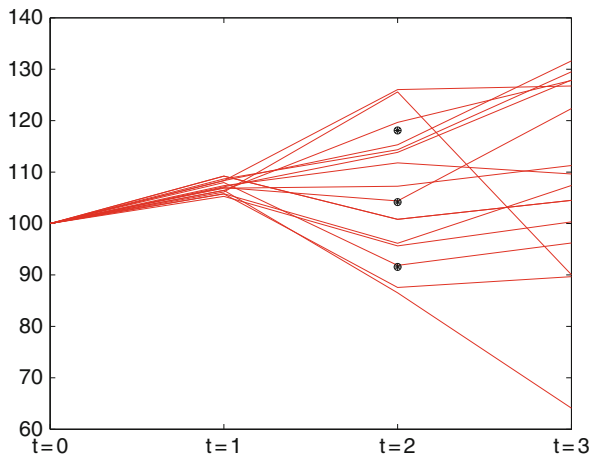
Step 2. Approximate the data in the first stage by minimal Kantorovich distance



Step 3. For the subsequent stages, use the paths which are close to its predecessor ...



Step 4. .. and iterate this step through the whole tree.



15.5.2 Methods Based on the Nested Distance

The general scenario tree problem is as follows:

15.5.2.1 The Scenario Tree Problem

Given a nested distribution \mathbb{P} of depth T , find a valuated tree (with nested distance $\tilde{\mathbb{P}}$) with at most S nodes such that the distance $d(\mathbb{P}, \tilde{\mathbb{P}})$ is minimal.

Unfortunately, this problem is such complex, that is, it is impossible to be solved for medium tree sizes. Therefore, some heuristic approximations are needed. A possible method is a stepwise and greedy reduction of a given tree. Two subtrees of the tree may be collapsed, if they are close in terms of the nested distance. Here is an example:

Example The three subtrees of level 2 of the tree in Fig. 15.9. We calculated their nested distances as follows:

$$d\left(\left[\begin{array}{c} \frac{1.0}{2.4} \\ \left[\frac{0.4}{5.1} \ 0.6\right] \end{array}\right], \left[\begin{array}{c} \frac{1.0}{3.0} \\ \left[\frac{1.0}{2.8}\right] \end{array}\right]\right) = 2.6,$$

$$d\left(\left[\begin{array}{c} \frac{1.0}{2.4} \\ \left[\frac{0.4}{5.1} \ 0.6\right] \end{array}\right], \left[\begin{array}{c} \frac{1.0}{3.0} \\ \left[\frac{0.4}{3.3} \ 0.2 \ 0.4\right] \end{array}\right]\right) = 2.62,$$

$$d\left(\left[\begin{array}{c} \frac{1.0}{2.4} \\ \left[\frac{0.4}{5.1} \ 0.6\right] \end{array}\right], \left[\begin{array}{c} \frac{1.0}{3.0} \\ \left[\frac{0.4}{3.3} \ 0.2 \ 0.4\right] \end{array}\right]\right) = 1.86.$$

To calculate a nested distance, we have to solve $\sum_{t=1}^T \#(N_t) \cdot \#(\tilde{N}_t)$ linear optimization problems, where $\#(N_t)$ is the number of nodes at stage t .

Example For the valued trees of Fig. 15.11 we get the following distances:

$$\begin{aligned} d(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}) &= 3.90; & d(P^{(1)}, P^{(2)}) &= 3.48, \\ d(\mathbb{P}^{(1)}, \mathbb{P}^{(3)}) &= 2.52; & d(P^{(1)}, P^{(3)}) &= 1.77, \\ d(\mathbb{P}^{(2)}, \mathbb{P}^{(3)}) &= 3.79; & d(P^{(2)}, P^{(3)}) &= 3.44. \end{aligned}$$

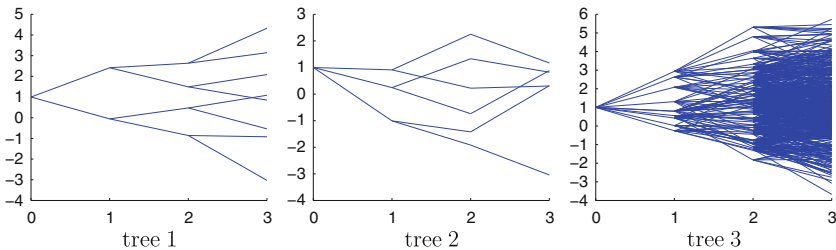


Fig. 15.11 Distances between nested distributions

Example Let \mathbb{P} be the (nested) probability distribution of the pair $\xi = (\xi_1, \xi_2)$ be distributed according to a bivariate normal distribution

$$\begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} = N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \right).$$

Note that the distribution of (ξ_1, ξ_2) is the same as the distribution of $(\zeta_1, \zeta_1 + \zeta_2)$, where ζ_i are independently standard normally distributed.

We know that the optimal approximation of a $N(\mu, \sigma^2)$ distribution in the Kantorovich sense by a three-point distribution is

$$\left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ \mu - 1.029\sigma & \mu & \mu + 1.029\sigma \end{array} \right].$$

The distance is 0.3397σ . Thus the optimal approximation to ξ_1 is

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

and the conditional distributions given these values are $N(-1.029, 1)$, $N(0, 1)$, and $N(1.029, 1)$. Thus the optimal two-stage approximation is

$$\tilde{\mathcal{P}}^{(1)} = \left[\begin{array}{c} \left[\begin{array}{ccc} 0.30345 & & 0.3931 & & 0.30345 \\ & -1.029 & & & 1.029 \end{array} \right] \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ -1.029 & 0.0 & 1.029 \end{array} \right] \left[\begin{array}{ccc} 0.30345 & 0.3931 & 0.30345 \\ 0.0 & 1.029 & 2.058 \end{array} \right] \end{array} \right].$$

The nested distance is $d(\mathbb{P}, \tilde{\mathcal{P}}^{(1)}) = 0.76$.

Figure 15.12 shows on the left the density of the considered normal model \mathbb{P} and on the right the discrete model $\tilde{\mathcal{P}}^{(1)}$.

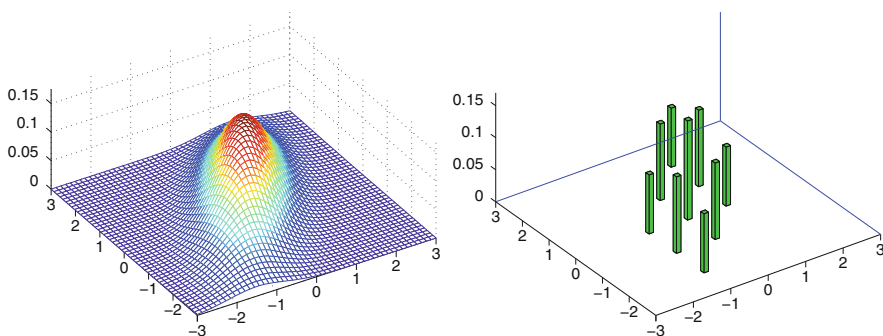


Fig. 15.12 Optimal two stage approximation

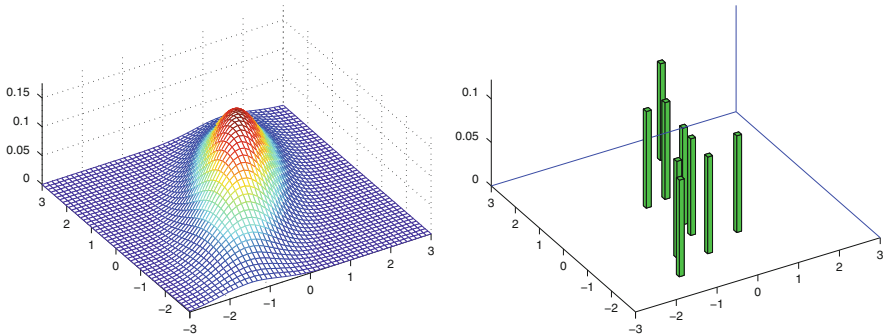


Fig. 15.13 Randomly sampled approximation

For comparison, we have also sampled nine random points from \mathbb{P} . Let $\tilde{\mathbb{P}}^{(2)}$ be the empirical distribution of these nine points. Figure 15.13 shows on the right side the randomly sampled discrete model $\tilde{\mathbb{P}}^{(2)}$. The nested distance is $\mathfrak{d}(\mathbb{P}, \tilde{\mathbb{P}}^{(2)}) = 1.12$.

15.6 Summary

In this contribution we have outlined some funding components to get stochastic optimization available for computational treatment. At first stage the quantization of probability measures is studied, which is essential to capture the initial problem from a computational perspective. Various distance concepts are available in general, however, only a few of them insure that the quality of the solution of the approximated problem is in line with the initial problem.

This concept then is extended to processes and trees. To quantify the quality of an approximating process, the concept of a nested distance is introduced. Some examples are given to illustrate this promising concept.

The collected ingredients give a directive, how to handle a stochastic optimization problem in general. Moreover – and this is at least as important – they allow to qualify solution statements of any optimization problem, which have stochastic elements incorporated.

References

- Z. Drezner and H.W. Hamacher. *Facility Location: Applications and Theory*. Springer, New York, NY, 2002.
- J. Dupacova. Stability and sensitivity analysis for stochastic programming. *Annals of Operations Research*, 27:115–142, 1990.
- Richard Durrett. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, second edition, 1996.
- A. Gibbs and F.E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.
- S. Graf and H. Luschgy. *Foundations of Quantization for Probability Distributions*. Lecture Notes in Math. 1730. Springer, Berlin, 2000.

- H. Heitsch and W. Roemisch. Scenario reduction algorithms in stochastic programming. *Computational Optimization and Applications*, 24:187–206, 2003.
- H. Heitsch, W. Roemisch, and C. Strugarek. Stability of multistage stochastic programs. *SIAM Journal on Optimization*, 17:511–525, 2006.
- C.C. Heyde. On a property of the lognormal distribution. *Journal of Royal Statistical Society Series B*, 25(2):392–393, 1963.
- R. Hochreiter and G. Ch. Pflug. Financial scenario generation for stochastic multi-stage decision processes as facility location problems. *Annals of Operation Research*, 152(1):257–272, 2007.
- K. Hoyland and S.W. Wallace. Generating scenario trees for multistage decision problems. *Management Science*, 47:295–307, 2001.
- S. Na and D. Neuhoff. Bennett’s integral for vector quantizers. *IEEE Transactions on Information Theory*, 41(4):886–900, 1995.
- H. Niederreiter. *Random Number Generation and Quasi-Monte Carlo Methods*, volume 23. CBMS-NSF Regional Conference Series in Applied Math., Soc. Industr. Applied Math., Philadelphia, PA, 1992.
- T. Pennanen. Epi-convergent discretizations of multistage stochastic programs via integration quadratures. *Mathematical Programming*, 116:461–479, 2009.
- G. Ch. Pflug. Scenario tree generation for multiperiod financial optimization by optimal discretization. *Mathematical Programming, Series B*, 89:251–257, 2001.
- G. Ch. Pflug. Version-independence and nested distributions in multistage stochastic optimization. *SIAM Journal on Optimization*, 20(3):1406–1420, 2009.
- S.T. Rachev and W. Roemisch. Quantitative stability in stochastic programming: The method of probability metrics. *Mathematics of Operations Research*, 27(4):792–818, 2002.
- S.T. Rachev. *Probability Metrics and the Stability of Stochastic Models*. Wiley, New York, NY, 1991.
- S.S. Vallander. Calculation of the Wasserstein distance between probability distributions on the line. *Theor. Prob. Appl.*, 18:784–786, 1973.
- C. Villani. *Optimal Transport*. Springer, Berlin, Heidelberg, 2008.
- A. Weber. *Theory of the Location of Industries*. The University of Chicago Press, Chicago, IL 1929.