Approximation and Ambiguity in stochastic optimization

Georg Ch. Pflug/B. Analui/F. Maggioni

June 2014
The Decision Problem

Multistage stochastic optimization problem

\[
\text{(Opt)} \quad \begin{array}{|c|}
\hline
\text{Minimize in } x_0, x_1, \ldots, x_{T-1} : \\
\mathcal{R}[Q(x_0, \xi_1, \ldots, x_{T-1}, \xi_T)] \\
\text{s.t. } x \triangleleft \mathcal{F} \\
\text{and possibly other constraints on } x_0, \ldots, x_{T-1} : x \in \mathcal{X} \\
\hline
\end{array}
\]

\[\mathcal{F} = \mathcal{F}_0, \ldots, \mathcal{F}_T\text{ is a filtraton. } \mathcal{X}\text{ is a convex set.}\]

\[x \triangleleft \mathcal{F}\text{ means that } x_t \triangleleft \mathcal{F}_t \text{ for all } t, \text{ i.e. that the decisions are nonanticipative.}\]
In order to numerically solve the multiperiod stochastic optimization problem, the stochastic process \((\xi_t)\) must be approximated by a simple stochastic process \(\tilde{\xi}_t\), which takes only a small number of values. Likewise the filtration \(\mathcal{F}\) must be approximated by a smaller one \(\tilde{\mathcal{F}}\) such that \(\sigma(\tilde{\xi}) \subseteq \tilde{\mathcal{F}}\).

\[
\begin{align*}
\text{Minimize in } & \tilde{x}_0, x_1, \ldots, \tilde{x}_{T-1} : \\
\mathcal{R}[Q(\tilde{x}_0, \tilde{\xi}_1, \ldots, \tilde{x}_{T-1}, \tilde{\xi}_T)] \\
\text{s.t. } & \tilde{x} \triangleleft \tilde{\mathcal{F}} \\
\text{and possibly other constraints } & \tilde{x} \in \tilde{X}.
\end{align*}
\]
Measuring closedness by the Wasserstein distance

\[ d_1(P, \tilde{P}) = \sup \{ \int h \, dP - \int h \, d\tilde{P} : |h(u) - h(v)| \leq d(u, v) \}. \]

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

\[ d_1(P, \tilde{P}) = \inf \{ \mathbb{E}(d(X, Y)) : (X, Y) \text{ is a bivariate r.v. with given marginal distributions } P \text{ and } \tilde{P} \}. \]

**Remark.** The bivariate measure \( \pi \) is called a transportation plan. If both measures sit on a finite number of mass points, then \( d_1(P, \tilde{P}) \) is the optimal value of an LP.
Interpretation as facility location/mass transportation problem
Discrete processes: trees and nested distributions

\[
\begin{pmatrix}
0.2 & 3.0 \\
0.4 & 0.2 & 0.4 \\
6.0 & 4.7 & 3.3
\end{pmatrix}
\begin{pmatrix}
0.3 \\
3.0 \\
2.8
\end{pmatrix}
\begin{pmatrix}
0.5 \\
2.4 \\
1.0
\end{pmatrix}
\]
A valuated tree

An exemplary finite tree representing the filtered probability space \((\Omega, \mathcal{F}, P)\). The scenario process \((\xi_t)\) sits on its nodes.
Distances between nested distributions

**Definition.** For two nested distributions \( \mathbb{P} \sim (\Xi, \mathcal{F}, P, \xi) \), \( \tilde{\mathbb{P}} \sim (\tilde{\Xi}, \tilde{\mathcal{F}}, \tilde{P}, \tilde{\xi}) \) and a distance function \( d \) on \( \mathbb{R}^m \) the *nested distance of order* \( r \geq 1 \) — denoted \( d_{l,r}(\mathbb{P}, \tilde{\mathbb{P}}) \) — is the optimal value of the optimization problem

\[
\begin{align*}
\text{minimize} \quad & \left( \int d \left( \xi(\omega), \tilde{\xi}(\tilde{\omega}) \right)^r \pi \left( d\omega, d\tilde{\omega} \right) \right)^{1/r} \\
\text{subject to} \quad & \pi \left( M \times \tilde{\Xi} \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right) = P \left( M \mid \mathcal{F}_t \right) \quad (M \in \mathcal{F}_T) \\
& \pi \left( \Xi \times N \mid \mathcal{F}_t \otimes \tilde{\mathcal{F}}_t \right) = \tilde{P} \left( N \mid \tilde{\mathcal{F}}_t \right) \quad (N \in \tilde{\mathcal{F}}_T)
\end{align*}
\]

where the infimum in is among all bivariate probability measures \( \pi \in \mathcal{P}(\Omega \times \Omega') \), which are measures on the product sigma algebra \( \mathcal{F}_T \otimes \tilde{\mathcal{F}}_T \).
Remark. Trees are recursive structures. A transportation plan $\pi$ between trees must be definable in a recursive way too. For two finite trees of same height, the nested distance can be defined by an LP. The dual representation of the nested distance is

$$d^r_{\pi} \left( \mathbb{P}, \tilde{\mathbb{P}} \right) = \sup M_0$$

subject to

$$M_t = M_0 + \sum_{s=1}^{t} (\lambda_s + \mu_s)$$

$$\mathbb{E}(\lambda_{s+1} | \mathcal{F}_s) = 0$$

$$\mathbb{E}(\mu_{s+1} | \tilde{\mathcal{F}}_s) = 0$$

$$M_T(\omega, \tilde{\omega}) \leq d(\xi(\omega), \tilde{\xi}(\tilde{\omega}))^r$$

The "optimal" process $M_t$ is a martingale w.r.t. the optimal transportation measure $\pi$:

$$d^r_{\pi} = \mathbb{E}_\pi(M_T) = M_0.$$
The main approximation result

Let $Q_L$ be the family of all real valued cost functions $Q(x_0, y_1, x_1, \ldots, x_{T-1}, y_T)$ 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 $(Opt(\mathbb{P}))$

$$v^*_Q(\mathbb{P}) := \min \{ \mathbb{E}_P[Q(x_0, \xi_1, x_1, \ldots, x_{T-1}, \xi_T)] : x \in \mathcal{X} \}.$$ 

and the approximative problem $(Opt(\tilde{\mathbb{P}}))$

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

**Theorem.** (A. Pichler and G.P., 2011) For $Q$ in $Q_L$

$$|v^*_Q(\mathbb{P}) - v^*_Q(\tilde{\mathbb{P}})| \leq L \cdot \text{d}_r(\mathbb{P}, \tilde{\mathbb{P}}).$$

This bound is sharp.
**Definition.** The functional $P \mapsto R_P(\cdot)$ is compound concave if the mapping $P \mapsto R_P(Y)$ is concave for all random variables $Y$ for which $R$ is defined.

**Example.** All distortion functionals of the form

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

are compound concave. Distortion functionals are e.g. the upper Average Value at Risk or the Gini-coefficient.

**Lemma.** If $R$ is compound concave, then $P \mapsto \nu^*(P)$ is also concave, where $P$ is the probability measure sitting on $\Omega = \mathcal{N}_T$, the leaves of the tree.
where $\Omega = \bigcup_i \Omega_i^{(j)}$ for all $j$ and

$$\Omega_i^{(j)} = \bigcup_{\Omega_s^{(j-1)} \subseteq \Omega_i^{(j)}} \Omega_s^{(j-1)}.$$
To a refinement chain there corresponds a chain of dissections of
the probability $P$ into probability measures $P_i^{(j)}$

\[
P
\]

\[
(P_1^{(\ell)}, \ldots, P_{m_\ell}^{(\ell)})
\]

\[
: \quad (P_1^{(2)}, \ldots, P_{m_2}^{(2)})
\]

\[
(P_1^{(1)} = \delta_{\omega_1}, \ldots, P_k^{(1)} = \delta_{\omega_k})
\]

such that

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

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

(iii) each $P_i^{(j)}$ can be written as a convex combination of
probabilities from the refined collection $\left\{ P_i^{(j-1)} \right\}$. 
It is evident that given such a refinement chain leads to a chain of lower bounds. Denoting

\[ v_j^* = \sum_{i=1}^{m_j} \pi_i^{(j)} v^*(P_i^{(j)}) \]

one gets

\[ v_1^* \leq v_2^* \leq \cdots \leq v_\ell^* \leq v^*(P). \]

Here \( v_1^* \) comes from the clairvoyant’s dissection \( \Omega = \bigcup_i \{\omega_i\} \), which is the finest and the relation \( v_1^* \leq v^*(P) \) is also known as wait-and-see \( \leq \) here-and-now.
Examples for Refinement Chains

A graphical representation of a refinement chain with just one scenario fixed. The collection \((\Omega_1^{(2)}, \ldots, \Omega_8^{(2)})\) is the "pair of scenarios" collection (with one of them fixed). Similarly the collection \((\Omega_1^{(3)}, \ldots, \Omega_4^{(3)})\) is the "triple of scenarios" collection.
A refinement chain with disjoint subsets.
Together with upper bounds (obtained by inserting subsolutions), lower bounds allow to find $\epsilon$-optimal solutions in cases where the solution of the main problem is too time consuming. Also, the subproblems may solved in parallel.

As an example, we took a multiperiod inventory control problem, the multistage generalization of the newsboy problem. The basic tree has 540 nodes and height 5.

We found lower bounds for several Refinement Chains with varying numbers of nodes per run.
Percentage deviation from the optimal objective value $v^*$ of a refinement chain with disjoint subsets for increasing values of complexity of calculation measured in CPU seconds.
Percentage deviation from the optimal objective value $v^*$ of refinement chain with one scenario fixed for increasing values of complexity of calculation measured in CPU seconds. The index $i$ of $v_i^*$ indicates the number of scenarios in one subproblem.
Percentage deviation from the optimal objective value $\nu^*$ of refinement chain with 8 scenarios fixed for increasing values of complexity of calculation measured in CPU seconds.
Traditionally, we do stochastic optimization in two steps:

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

According to Ellsberg (1961) we face here two types of non-determinism:

*Uncertainty*: the probabilistic model is known, but the realizations of the random variables are unknown ("aleatoric uncertainty")

*Ambiguity*: the probability model itself is not fully known ("epistemic uncertainty" - Knightian uncertainty according to F. Knight "Risk, Uncertainty and Profit" (1920)).

Ambiguity sets: A family of probability models \( \mathcal{P} \) which are all plausible models for the reality and we are uncertain about which concrete \( P \in \mathcal{P} \) is the true one.
Let the basic problem be

\[
\min \{ F_{\hat{P}}(x) = R_{\hat{P}}[Q(x, \xi)] : x \in X \}
\]

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

\[
\min \{ F_{\mathcal{P}}(x) = \max \{ R_{\mathcal{P}}[Q(x, \xi)] : \mathbb{P} \in \mathcal{P} \} : x \in X \}.
\]

It is a minimax problem. A solution to this problem is also called *distributionally robust*. 
Shapiro and Kleywegt (2002) define a set of probability measures \( \mu \) such that

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

Shapiro and Ahmed (2004) define the ambiguity set as

\[
P = \left\{ \mu \text{ is a prob. measure s.t. } \mu_1 \prec \mu \prec \mu_2, \int \phi_i \, d\mu = b_i; i = 1, \ldots k; \int \psi \, d\mu \leq c_i; i = 1, \ldots \ell \right\}
\]

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

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

Delage and Ye (2010) consider the following ambiguity set

\[ \mathcal{P} = \left\{ \mu : \mu(S) = 1, (\int x \, d\mu(x) - c)^\top \Sigma_0^{-1} (\int x \, d\mu(x) - c) \leq \gamma_1, \right. \]
\[ \left. \int (x - c)(x - c)^\top d\mu(x) \leq \gamma_2 \Sigma_0 \right\} \]

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.
Edirisinghe (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.
Nested balls as ambiguity sets

Let

\[ \mathcal{P}_\epsilon = \{ \mathbb{P} : d_l(\hat{\mathbb{P}}, \mathbb{P}) \leq \epsilon \}. \]

We restrict ourselves in the following to alternative models, which are defined on the same tree structure \( T \) but only vary the probabilities. The alternative nested distributions are written as \( \mathbb{P}(T, P) \), \( P \) a probability on \( \Omega = \mathcal{N}_T \).

The final formulation of the ambiguity extension problem is now

\[
\min_x \max_{\mathbb{P} \in \mathcal{P}_\epsilon} \{ \mathcal{R}_\mathbb{P}[Q(x, \xi)] : x \in X, x \prec \mathcal{F} \}.
\]

where

\[ \mathcal{P}_\epsilon = \{ \mathbb{P}(T, P) : d_l(\mathbb{P}(T, P), \mathbb{P}(T, \hat{P})) \leq \epsilon \} \]
A minimax Theorem

The correct way of forming convex combinations of processes sitting on trees is by compounding.

\[ C(\mathbb{P}, \hat{\mathbb{P}}; \lambda) = \begin{cases} \mathbb{P} & \text{with prob } \lambda \\ \hat{\mathbb{P}} & \text{with prob } 1 - \lambda \end{cases} \]

Nested balls are not convex w.r.t. compounding. So, let \( \bar{\mathcal{P}}_\epsilon \) be the closed convex hull of the ambiguity set \( \mathcal{P}_\epsilon \).

**Theorem.** Let \( Q(x, \xi) \) be convex in \( x \) with a convex and compact decision set \( X \). Then

\[
\min_{x \in X} \max_{\mathbb{P} \in \bar{\mathcal{P}}_\epsilon} \mathbb{R}_\mathbb{P}[Q(x, \xi)] = \max_{\mathbb{P} \in \bar{\mathcal{P}}_\epsilon} \min_{x \in X} \mathbb{R}_\mathbb{P}[Q(x, \xi)]
\]

and there exists a saddle point \( (x^*, \mathbb{P}^*) \), i.e.

\[
\mathbb{R}_\mathbb{P}[H(x^*, \xi)] \leq \mathbb{R}_{\mathbb{P}^*}[Q(x^*, \xi)] \leq \mathbb{R}_{\mathbb{P}^*}[Q(x, \xi)]
\]

for all \( x \in X, \mathbb{P} \in \bar{\mathcal{P}}_\epsilon \). Moreover, \( \mathbb{P}^* \) can be chosen to lie in \( \mathcal{P}_\epsilon \) (and not just in \( \bar{\mathcal{P}}_\epsilon \)).
The price of ambiguity and the gain of robustness

Let $x^*(\hat{P})$ be the solution of the baseline problem and let $x^*(P)$ be the minimax solution. Then one may define the following nonnegative quantities

**Gain of robustness.**

$$\max_{\mathcal{P} \in \mathcal{P}} \mathcal{R}_{\hat{P}}[Q(x^*(\hat{P}), \xi)] - \max_{\mathcal{P} \in \mathcal{P}} \mathcal{R}_P[Q(x^*(P), \xi)].$$

**Price for ambiguity.**

$$\mathcal{R}_{\hat{P}}[Q(x^*(P), \xi)] - \mathcal{R}_\hat{P}[Q(x^*(\hat{P}), \xi)].$$
Successive convex programming (SCP)

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

$$\begin{array}{ll}
\min & u \\
\text{s.t.} & \mathcal{R}_P[Q(x, \xi)] \leq u \quad \text{for all} \quad P \in \mathcal{P}_\epsilon^k \\
& x \in X, \\
& x \in \mathcal{F}
\end{array}$$

and call the solution $(x_k, t_k)$.

3. Solve the inner problem

$$\begin{array}{ll}
\max & \mathcal{R}_P[Q(x^k, \xi)] \\
\text{s.t.} & P \in \mathcal{P}_\epsilon^k
\end{array}$$

Call the solution $P^k$ and let $\mathcal{P}_\epsilon^{k+1} = \mathcal{P}_\epsilon^k \cup \{P^k\}$. This optimization is typically approximated by successive linear programming.

4. Stop if no change in the minimax value is observed.
The scenario process consists of 5 components: Spot prices, Pumping prices, Inflows for 3 reservoirs. We construct a tree of height 8 with 421 nodes and 119 leaves.
The typical picture: The larger is the ambiguity radius, the simpler is the worst case tree.
The worst case 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%.
Gain of robustness: 7.5%.
“Multistage Stochastic Optimization”
by Alois Pichler and Georg Pflug
Springer Series in Operations Research and Financial Engineering
in print