Annals of Operations Research manuscript No.
(will be inserted by the editor)

Scenario decomposition of risk-averse multistage stochastic programming problems

Ricardo A. Collado · Dávid Papp · Andrzej Ruszczyński

This paper is dedicated to András Prékopa on the occasion of his 80th birthday

Received: date / Accepted: date

Abstract For a risk-averse multistage stochastic optimization problem with a finite scenario tree, we introduce a new scenario decomposition method and we prove its convergence. The main idea of the method is to construct a family of risk-neutral approximations of the problem. The method is applied to a risk-averse inventory and assembly problem. In addition, we develop a partially regularized bundle method for nonsmooth optimization.

Keywords dynamic measures of risk · duality · decomposition · bundle methods

1 Introduction

In the last decade the theory of coherent risk measures established itself as an alternative to expected utility models of risk averse preferences in stochastic optimization. This theory was initiated in [1,3] and further developed in numerous publications (see, e.g., [9,10,21,26,27] and the references therein). Recently, increased attention is paid to dynamic measures of risk, which allow for risk-averse evaluation of streams of future costs or rewards (see, e.g., [2,7,11,18,20,26,28,30]).

When used in stochastic optimization models, dynamic risk measures lead to a new class of problems, which are significantly more complex than their risk-neutral counterparts (see [27–29,31]). Decomposition, an established and efficient approach

Ricardo A. Collado
RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway, NJ 08854, USA
E-mail: collador@eden.rutgers.edu

Dávid Papp
RUTCOR, Rutgers University, 640 Bartholomew Road, Piscataway, NJ 08854, USA
E-mail: dpapp@rutcor.rutgers.edu

Andrzej Ruszczyński
Department of Management Science and Information Systems, 94 Rockefeller Road, Rutgers University, Piscataway, NJ 08854, USA
E-mail: rusz@business.rutgers.edu

to risk-neutral multistage stochastic optimization problems (see [6, 13, 19, 23] and the references therein), cannot be directly applied to risk-averse models. With dynamic risk measures, the main feature facilitating decomposition, the integral form of the objective function, is absent. Our main objective is to overcome this difficulty by exploiting specific structure of dynamic risk measures, and to develop new decomposition methods that extend the ideas of earlier approaches to risk-neutral problems. We initiated this research in [16], where we developed risk-averse counterparts of the primal (Benders-type) decomposition methods.

In this paper we develop generalizations of scenario decomposition methods, in the spirit of [17]. The key to success is the use of dual properties of dynamic measures of risk to construct a family of risk-neutral approximations of the problem. In sections 2 and 3 we formally define a multistage risk-averse stochastic optimization problem and we discuss its properties. Section 4 discusses nonanticipativity constraints. In section 5 we advance the duality theory of dynamic measures of risk, by identifying the properties that are essential for our decomposition approach. In section 6 we present the main idea of our new decomposition methods. In section 7 we analyze properties of the master (coordination) problem of the method. Finally, section 8 is devoted to the application of two versions of our methods, with several coordination algorithms, to an inventory planning and assembly problem. In the development of efficient master algorithms we modify the bundle method, to better exploit the specificity of the problem at hand. The resulting algorithm, which we call the partial bundle method, is discussed in the appendix.

2 A Multistage Risk-Averse Problem

Let $(\Omega, \mathscr{F}, P)$ be a probability space with a sigma algebra $\mathscr{F}$ and probability measure P. Consider a filtration $\{\emptyset, \Omega\} = \mathscr{F}_1 \subset \mathscr{F}_2 \subset \cdots \subset \mathscr{F}_T = \mathscr{F}$. A random vector $x = (x_1, \ldots, x_T)$, where each x_t has values in $\mathbb{R}^{n_t}$, $t = 1, \ldots, T$, is called a *policy*. If each x_t is $\mathscr{F}_t$-measurable, $t = 1, \ldots, T$, a policy x is called *implementable*. A policy x is called *feasible*, if it satisfies the following conditions:

$$\begin{array}{rcccccl} A_1 x_1 & & & & & & = b_1, \\ B_2 x_1 & + & A_2 x_2 & & & & = b_2, \\ & & B_3 x_2 & + & A_3 x_3 & & = b_3, \\ \multicolumn{6}{c}{\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots} \\ & & & & B_T x_{T-1} + & A_T x_T & = b_T, \\ \multicolumn{6}{l}{x_1 \in X_1, \quad x_2 \in X_2, \quad x_3 \in X_3, \quad \ldots \quad x_T \in X_T.} \end{array} \tag{1}$$

In these equations, for every $t = 1, \ldots, T$, the matrices A_t of dimensions $m_t \times n_t$, the matrices B_t of dimensions $m_t \times n_{t-1}$, and the vectors b_t of dimensions m_t are $\mathscr{F}_t$-measurable data. Each set X_t is a random convex and closed polyhedron which is measurable with respect to $\mathscr{F}_t$ (in the sense of measurability of multifunctions; see [4]).

Suppose c_t, $t = 1, \ldots, T$, is an adapted sequence of random cost vectors, that is, each c_t is $\mathscr{F}_t$-measurable. A policy x results in a cost sequence

$$Z_t = \langle c_t, x_t \rangle, \quad t = 1, \ldots, T. \tag{2}$$

Our intention is to formulate and analyze a risk-averse multistage stochastic programming problem, to minimize a dynamic measure of risk, $\rho(Z_1,\dots,Z_T)$, over all implementable and feasible policies x. In order to define the functional $\rho(\cdot)$, we recall some basic concepts of the theory of dynamic measures of risk. We follow the development given in [27–29,31].

Consider vector spaces $\mathscr{Z}_t$ of $\mathscr{F}_t$-measurable random outcomes. As $\mathscr{F}_1$ is trivial, $\mathscr{Z}_1 = \mathbb{R}$. For $Z, Z' \in \mathscr{Z}_T$ we denote by $Z \preceq Z'$ the pointwise partial order meaning $Z_t(\omega) \le Z'_t(\omega)$ for a.e. $\omega \in \Omega$.

Let $1 \le t \le T-1$. A *coherent conditional risk measure* is a function $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$ satisfying the following axioms:

(A1) *Convexity*: $\rho_t(\alpha Z + (1-\alpha)Z') \preceq \alpha\rho_t(Z) + (1-\alpha)\rho_t(Z')$, for all $Z, Z' \in \mathscr{Z}_{t+1}$ and all $\alpha \in [0,1]$;
(A2) *Monotonicity*: If $Z, Z' \in \mathscr{Z}_{t+1}$ and $Z \preceq Z'$, then $\rho_t(Z) \preceq \rho_t(Z')$;
(A3) *Predictable Translation Equivariance*: If $V \in \mathscr{Z}_t$ and $Z \in \mathscr{Z}_{t+1}$, then $\rho_t(V + Z) = V + \rho_t(Z)$;
(A4) *Positive Homogeneity*: If $\gamma \ge 0$ and $Z \in \mathscr{Z}_{t+1}$, then $\rho(\gamma Z) = \gamma\rho(Z)$.

We assume that the smaller the realizations of Z, the better; for example Z may represent a random cost.

An example of coherent conditional risk measure is the *conditional mean–upper semideviation* model defined by

$$\rho_t(Z) = \mathbb{E}[Z|\mathscr{F}_t] + \kappa_t \,\mathbb{E}\left[\left(Z - \mathbb{E}[Z|\mathscr{F}_t]\right)_+ \big|\mathscr{F}_t\right], \tag{3}$$

with an $\mathscr{F}_t$-measurable $\kappa_t \in [0,1]$. See [31, page 277] for the details showing that the mean upper semideviation is a coherent conditional risk measure, and for other examples of conditional risk measures.

Suppose we observe a random sequence Z_t, $t = 1,\dots,T$, adapted to the filtration $\{\mathscr{F}_t\}$. Its risk can be evaluated by using the following *dynamic coherent measure of risk*

$$\rho_{1,T}(Z_1, Z_2, \dots, Z_T) = Z_1 + \rho_1\Big(Z_2 + \rho_2\big(Z_3 + \cdots + \rho_{T-1}(Z_T)\dots\big)\Big), \tag{4}$$

where each $\rho_t : \mathscr{Z}_{t+1} \to \mathscr{Z}_t$ is a coherent conditional measure of risk. The structure (**??**) was postulated in [28] and derived in [25] from abstract principles of monotonicity and time consistency of dynamic risk measures.

Our problem is to minimize (4) with each Z_t given by (2), over all implementable and feasible policies x. In order to complete the problem formulation, we need to be more specific about the vector spaces $\mathscr{Z}_t$, the vector spaces of random vectors in which the components x_t of the policy live, as well as integrability conditions on the problem data A_t, B_t, b_t and c_t, so that $Z_t \in \mathscr{Z}_t$ for all $t = 1,\dots,T$. In this paper, we assume that all sigma-algebras are finite and all vector spaces are finite-dimensional. We discuss it in the next section.

3 Scenario Trees and Recursive Risk Evaluation

In the finite distribution case, possible realizations of data form a *scenario tree*. It has nodes organized in levels which correspond to stages $1,\dots,T$. At level $t=1$ we have only one *root node* $\nu=1$. Nodes at levels $t=2,\dots,T$ correspond to elementary events in $\mathscr{F}_t$. Each node ν at level $t=2,\dots,T$ is connected to a unique node $a(\nu)$ at level $t-1$, called the *ancestor node*, which corresponds to the elementary event in $\mathscr{F}_{t-1}$ that contains the event associated with ν. Thus, every node ν at levels $t=1,\dots,T-1$ is connected to a set $C(\nu)$ of nodes at level $t+1$, called *children nodes*, which correspond to elementary events in $\mathscr{F}_{t+1}$ included in the event corresponding to ν. We denote by Ω_t the set of all nodes at stage $t=1,\dots,T$. We have the relations $\Omega_{t+1}=\cup_{\nu\in\Omega_t}C(\nu)$ and $C(\nu)=\{\eta\in\Omega_{t+1}:\nu=a(\eta)\}$. The sets $C(\nu)$ are disjoint, i.e., $C(\nu)\cap C(\nu')=\emptyset$ if $\nu\neq\nu'$. A *scenario* is a path s from the root to a node at the last stage T. By construction, there is one-to-one correspondence between the scenarios and the set $\Omega_T=\Omega$. Let $\mathscr{S}(\nu)$ be the set of scenarios passing through node ν. These sets satisfy the recursive relation:

$$\begin{aligned}\mathscr{S}(\nu)&=\{\nu\},\quad \nu\in\Omega_T,\\ \mathscr{S}(\nu)&=\bigcup_{\eta\in C(\nu)}\mathscr{S}(\eta),\quad \nu\in\Omega_t,\quad t=T-1,\dots,1.\end{aligned}$$

As the nodes of the tree correspond to events defining nested partitions of Ω, the measure P can be specified by conditional probabilities:

$$p_{\nu\eta}=P[\eta|\nu],\quad \nu\in\Omega_t,\quad \eta\in C(\nu),\quad t=1,\dots,T-1.$$

Every node ν at level t has a *history*: the path $(\nu_1,\dots,\nu_{t-1},\nu)$ from the root to ν. The probability of the node ν is thus the product of the corresponding conditional probabilities

$$p_\nu=p_{\nu_1\nu_2}p_{\nu_2\nu_3}\cdots p_{\nu_{t-1}\nu}. \tag{5}$$

In particular, when $t=T$, formula (5) describes the probability of a scenario $\nu\in\Omega_T$.

For every node $\nu\in\Omega_t$, an $\mathscr{F}_t$-measurable random variable Z has identical values on all scenarios $s\in\mathscr{S}(\nu)$. It can, therefore, be equivalently represented as a function of a node at level Ω_t, which we write Z^{Ω_t}.

Consider a conditional measure of risk $\rho_t(\cdot)$. Its value is $\mathscr{F}_t$-measurable, and thus we can consider its representation as a function of a node at level t. It follows from [28, Thm. 3.2] that for every $\mathscr{F}_t$-measurable nonnegative function Γ a stronger version of (A4) holds:

$$\Gamma\rho_t(Z_{t+1})=\rho_t(\Gamma Z_{t+1}).$$

Let $\nu\in\Omega_t$, and let $\mathbb{1}_\nu$ be the characteristic function of the event ν. Setting $\Gamma=\mathbb{1}_\nu$ in the last equation, for all $Z_{t+1},W_{t+1}\in\mathscr{Z}_{t+1}$ we obtain

$$\mathbb{1}_\nu\rho_t(\mathbb{1}_\nu Z_{t+1}+(1-\mathbb{1}_\nu)W_{t+1})=\rho_t(\mathbb{1}_\nu Z_{t+1})=\mathbb{1}_\nu\rho_t(\mathbb{1}_\nu Z_{t+1}).$$

In the last equation we multiplied both sides by $\mathbb{1}_\nu$. We see that W_{t+1} plays no role here. The value of $\rho_t(Z_{t+1})$ at elementary events associated with node ν depends only

on the values of $Z_{t+1}^{\Omega_{t+1}}$ at nodes $\eta \in C(\nu)$. We denote the vector of these values by $Z_{t+1}^{C(\nu)}$, and we write the conditional risk measure equivalently as $\rho_t^{\nu}\big(Z_{t+1}^{C(\nu)}\big)$.

Let us define the random variables

$$V_t = \rho_t\Big(Z_{t+1} + \rho_{t+1}\big(Z_{t+2} + \cdots + \rho_{T-1}(Z_T)\ldots\big)\Big), \quad t = 1,\ldots,T. \tag{6}$$

They are $\mathscr{F}_t$-measurable, and thus we only need to consider their values V_t^{ν} associated with scenarios $s \in \mathscr{S}(\nu)$. It follows that the value of the measure of risk (4) can be written on the scenario tree in a recursive manner:

$$\rho_{1,T}(Z_1, Z_2, \ldots, Z_T) = Z_1 + V_1^1, \tag{7}$$

$$V_t^{\nu} = \rho_t^{\nu}\big(Z_{t+1}^{C(\nu)} + V_{t+1}^{C(\nu)}\big), \quad \nu \in \Omega_t, \quad t = 1,\ldots,T. \tag{8}$$

4 Nonanticipativity Constraints

A standard approach to multistage stochastic programming is based on *scenario decomposition*. With every scenario s in the tree, we associate a sequence of decision vectors

$$x^s = (x_1^s, \ldots, x_T^s), \quad s \in \Omega.$$

Such a collection of sequences forms a policy which is not necessarily implementable, unless it satisfies a certain linear equation, called the *nonanticipativity constraint*. It requires that the process x be adapted to the filtration $\{\mathscr{F}_t\}$. Abstractly, we can write

$$x_t = \mathbb{E}\big[x_t | \mathscr{F}_t\big], \quad t = 1,\ldots,T-1. \tag{9}$$

For the scenario model, the nonanticipativity constraint can be written as a system of linear equations at the nodes of the tree. For every node ν at level $t = 1,\ldots,T-1$ the values x_t^s should be identical for all $s \in \mathscr{S}(\nu)$. Direct specification of (9) yields

$$x_t^s = \mathbb{E}\big[x_t | \mathscr{S}(\nu)\big] = \frac{\sum_{\omega \in \mathscr{S}(\nu)} p_\omega x_t^\omega}{\sum_{\omega \in \mathscr{S}(\nu)} p_\omega}, \quad s \in \mathscr{S}(\nu), \quad \nu \in \Omega_t, \quad t = 1,\ldots,T-1. \tag{10}$$

Other constraints of problem (1)-(2)-(3) decompose by scenario:

$$x \in \mathscr{X} = \mathscr{X}^1 \times \cdots \times \mathscr{X}^{|\Omega|}, \tag{11}$$

where for each $s \in \Omega$ we have

$$\mathscr{X}^s = \big\{x \in X_1^s \times \cdots \times X_T^s : B_t^s x_{t-1}^s + A_t^s x_t^s = b_t^s,\ t = 1,\ldots,T\big\}. \tag{12}$$

In (12) the symbols A_t^s, B_t^s, b_t^s, and X_t^s denote realizations of problem data at stage t in scenario s, and the term $B_t^s x_{t-1}^s$ is omitted for $t = 1$.

In risk-neutral multistage stochastic programming, we can write the corresponding optimization problem:

$$\begin{aligned} &\min \sum_{s\in\Omega} p^s \sum_{t=1}^{T} \langle c_t^s, x_t^s \rangle \\ &\text{s.t. (10) and (12).} \end{aligned} \tag{13}$$

Then, Lagrange multipliers λ_t^s are associated with the nonanticipativity constraints (10), and the following Lagrangian function is constructed:

$$L(x,\lambda) = \sum_{s\in\Omega} p^s \sum_{t=1}^{T} \langle c_t^s, x_t^s\rangle + \sum_{t=1}^{T-1} \sum_{\nu\in\Omega_t} \sum_{s\in\mathscr{S}(\nu)} p^s \big\langle \lambda_t^s, x_t^s - \mathbb{E}\big[x_t|\mathscr{S}(\nu)\big]\big\rangle. \tag{14}$$

The problem

$$\min_{x\in\mathscr{X}} L(x,\lambda)$$

decomposes into scenario subproblems, one for each $s\in\Omega$. We shall not go into these details here; the reader can find them in [31, Sec. 3.2.4]. The dual problem is to find the optimal values of Lagrange multipliers associated with (10). It can be solved by nonsmooth optimization methods or by augmented Lagrangian methods. As the constraints (10) are redundant, we can restrict the multipliers to the subspace defined by the equations

$$\mathbb{E}[\lambda_t|\mathscr{F}_t] = 0, \quad t = 1,\dots,T-1. \tag{15}$$

In the scenario tree case, these conditions translate into

$$\sum_{s\in\mathscr{S}(\nu)} p^s \lambda_t^s = 0, \quad \nu\in\Omega_t, \quad t = 1,\dots,T-1. \tag{16}$$

Again, the reader is referred to [31, Ch. 3] for the details.

The difficulty with the scenario decomposition in the risk-averse setting is the definition and nonlinear character of the dynamic risk measure (4). If a policy x is not implementable, the sequence $\{Z_t\}$ is not adapted to the filtration $\{\mathscr{F}_t\}$ and formula (4) makes no sense, because of the definition of ρ_t as a function acting on $\mathscr{F}_{t+1}$-measurable random variables. We cannot just substitute the dynamic risk measure for the objective function in (13).

5 Transition Multikernels and Their Compositions

We first recall the dual representation of conditional measures of risk. Let $\mathscr{P}(C)$ denote the set of probability distributions on a set of nodes $C\subset\Omega_t$. By [28, Remark 4.3], for every $t = 1,\dots,T-1$ and every node $\nu\in\Omega_t$ there exists a convex closed set $\mathscr{A}_t(\nu)\subset\mathscr{P}(C(\nu))$ such that

$$\rho_t^\nu\big(Z_{t+1}^{C(\nu)}\big) = \max_{\mu\in\mathscr{A}_t(\nu)} \big\langle \mu, Z_{t+1}^{C(\nu)}\big\rangle. \tag{17}$$

In fact, $\mathscr{A}_t(\nu) = \partial\rho_t^\nu\big(0\big)$.

We shall call a mapping $\mathscr{K}:\Omega_t\rightrightarrows\mathscr{P}(\Omega_{t+1})$ a *transition multikernel*. It is convex, if for all $\nu\in\Omega_t$ the set $\mathscr{K}(\nu)$ is convex. It is closed, if for all $\nu\in\Omega_t$ the set $\mathscr{K}(\nu)$ is closed. The transition multikernels $\mathscr{A}_t$ associated with the conditional risk measures $\rho_t(\cdot)$ are convex and closed, as subdifferentials of convex functions $\rho_t^\nu(\cdot)$ at 0, $\nu\in\Omega_t$. They also satisfy the conditions

$$\mathscr{A}_t(\nu)\subset\mathscr{P}(C(\nu)), \quad \forall\,\nu\in\Omega_t. \tag{18}$$

For $t = 1$ there is only one node $\nu = 1 \in \Omega_1$, and thus $\mathscr{A}_1$ is simply a set probability distributions on Ω_2. If a kernel μ_t is a selection of $\mathscr{A}_t$, that is, $\mu_t(\nu) \in \mathscr{A}_t(\nu)$ for all $\nu \in \Omega_t$, we shall simply write $\mu_t \in \mathscr{A}_t$. The value of $\mu(\nu)$ at an node $\eta \in C(\nu)$ will be written as $\mu(\nu, \eta)$.

Compositions of transition multikernels are germane for our analysis. Let us start from a composition of a measure $q_t \in \mathscr{P}(\Omega_t)$ with a kernel $\mu_t \in \mathscr{A}_t$ It is a measure on Ω_{t+1} given by the following relations:

$$(\mu_t \circ q_t)(\eta) = q_t(a(\eta))\mu_t(a(\eta), \eta), \ \eta \in \Omega_{t+1}; \tag{19}$$

recall that $a(\eta)$ is the ancestor of η. If we have a set of probability distributions $\mathscr{Q}_t \subset \mathscr{P}(\Omega_t)$ and a transition multikernel $\mathscr{A}_t$ satisfying (18), we can define their composition $\mathscr{A}_t \circ Q_t$ as the following set of probability distributions on Ω_{t+1}:

$$\mathscr{A}_t \circ Q_t = \big\{\mu_t \circ q_t : q_t \in Q_t, \mu_t \in \mathscr{A}_t\big\}. \tag{20}$$

Lemma 1 *Suppose Q_t is a convex and compact set of probability measures on Ω_t and a transition multikernel $\mathscr{A}_t$ satisfies* (18) *and is convex and compact. Then the set $Q_{t+1} = \mathscr{A}_t \circ Q_t$ is convex and compact.*

Proof To prove convexity, let $q^k_{t+1}(\eta) = q^k_t(a(\eta))\mu^k_t(\eta)$, with $\mu^k_t \in \mathscr{A}_t$, $q^k_t \in Q_t$, $k = 1, 2$, and consider their convex combination,

$$q_{t+1} = \alpha q^1_{t+1} + (1-\alpha)q^2_{t+1}, \quad \alpha \in (0,1).$$

Define $q_t = \alpha q^1_t + (1-\alpha)q^2_t$. By the convexity of Q_t, we have $q_t \in Q_t$, and thus the set $\mathscr{A}_t \circ \{q_t\}$ is included in Q_{t+1}. To show that $q_{t+1} \in Q_{t+1}$, it is sufficient to prove that $q_{t+1} \in \mathscr{A}_t \circ \{q_t\}$. This amounts to verifying for all $\eta \in \Omega_{t+1}$ the following relation:

$$\alpha q^1_t(a(\eta))\mu^1_t(\eta) + (1-\alpha)q^2_t(a(\eta))\mu^2_t(\eta) \in q_t(a(\eta))\mathscr{A}_t(a(\eta)). \tag{21}$$

Let $\eta \in \Omega_{t+1}$ and $\nu = a(\eta)$. Observe that $q^1_t(\nu) \geq 0$ and $q^2_t(\nu) \geq 0$. If $q_t(\nu) = 0$, we must have $q^1_t(\nu) = q^2_t(\nu) = 0$ and (21) is trivial. It remains to consider the case of $q_t(\nu) > 0$. Define

$$\beta(\nu) = \frac{\alpha q^1_t(\nu)}{q_t(\nu)}.$$

By the definition of q_t, $\beta(\nu) \in [0,1]$. The left hand side of (21) can be written as follows

$$\alpha q^1_t(\nu)\mu^1_t(\eta) + (1-\alpha)q^2_t(\nu)\mu^2_t(\eta) = q_t(\nu)\big(\beta(\nu)\mu^1_t(\eta) + (1-\beta(\nu))\mu^2_t(\eta)\big).$$

Due to the convexity of $\mathscr{A}_t$, the right hand side is an element of $\mathscr{A}_t \circ \{q_t\}$, which proves (21).

The compactness of Q_{t+1} follows from the compactness of Q_t and $\mathscr{A}_t$. □

We can now prove a useful dual representation of a dynamic measure of risk.

Theorem 1 *Suppose a dynamic risk measure $\rho(\cdot)$ is given by* (4) *with conditional risk measures $\rho_t(\cdot)$ satisfying conditions* (A1)–(A4)*. Then for every adapted sequence $Z_1,\dots,Z_T$ we have the relation*

$$\rho(Z_1,\dots,Z_T) = \max_{q_T\in Q_T} \big\langle q_T, Z_1+Z_2+\cdots+Z_T\big\rangle, \tag{22}$$

where

$$Q_T = \mathscr{A}_{T-1}\circ\ldots\mathscr{A}_2\circ\mathscr{A}_1 \tag{23}$$

is a convex and closed set of probability measures on Ω.

Proof Recursive composition of transition multikernels μ_t yields a sequence of sets of measures:

$$Q_{t+1} = \mathscr{A}_t\circ Q_t, \quad t=1,\dots,T-1, \tag{24}$$

with $Q_1=\{1\}$. Each Q_t is a set of probability measures on Ω_t. Lemma 1 implies that they are all convex and compact.

The multikernel representation (17) allows us to rewrite the definition of a dynamic risk measure (4) as follows:

$$\begin{aligned}\rho(Z_1,\dots,Z_T) = Z_1 + \max_{\mu_1\in\mathscr{A}_1}\Big(\big\langle\mu_1, Z_2^{\Omega_2}\big\rangle + \max_{\mu_2\in\mathscr{A}_2}\big(\big\langle\mu_2\circ\mu_1, Z_3^{\Omega_3}\big\rangle+\dots\\ \cdots+\max_{\mu_{T-1}\in\mathscr{A}_{T-1}}\big\langle\mu_{T-1}\circ\cdots\circ\mu_2\circ\mu_1, Z_T\big\rangle\cdots\big)\Big).\end{aligned} \tag{25}$$

All the maximum operations can be put at the beginning, and we obtain:

$$\begin{aligned}\rho(Z_1,\dots,Z_T) = Z_1 + \max_{\substack{\mu_t\in\mathscr{A}_t\\ t=1,\dots,T-1}}\Big(\big\langle\mu_1, Z_2^{\Omega_2}\big\rangle + \big\langle\mu_2\circ\mu_1, Z_3^{\Omega_3}\big\rangle+\dots\\ \cdots+\big\langle\mu_{T-1}\circ\cdots\circ\mu_2\circ\mu_1, Z_T\big\rangle\Big).\end{aligned} \tag{26}$$

Let $q_t=\mu_{t-1}\circ\cdots\circ\mu_2\circ\mu_1$, $t=2,\dots,T$. Each of them is an element of the corresponding set Q_t. Consider the product

$$\big\langle q_t, Z_t^{\Omega_t}\big\rangle = \sum_{\nu\in\Omega_t} q_t(\nu) Z_t^{\Omega_t}(\nu).$$

Suppose $\mu_t\in\mathscr{A}_t$ and $\nu\in\Omega_t$. Then $\mu_t(\nu)$ is a probability distribution on $C(\nu)$. Since Z_t is $\mathscr{F}_t$-measurable, $Z_t^{\Omega_{t+1}}$ has identical values on the nodes $\eta\in C(\nu)$. Therefore,

$$Z_t^{\Omega_t}(\nu) = \langle\mu_t(\nu), Z_t^{\Omega_{t+1}}\rangle.$$

Recalling the definition (19), we conclude that

$$\big\langle q_t, Z_t^{\Omega_t}\big\rangle = \big\langle\mu_t\circ q_t, Z_t^{\Omega_{t+1}}\big\rangle = \big\langle q_{t+1}, Z_t^{\Omega_{t+1}}\big\rangle.$$

Applying this relation recursively to all terms of (26), we obtain the identity

$$\begin{aligned}\rho(Z_1,\dots,Z_T) &= \max_{\substack{\mu_t\in\mathscr{A}_t\\ t=1,\dots,T-1}}\big\langle\mu_{T-1}\circ\cdots\circ\mu_2\circ\mu_1, Z_1+Z_2+\cdots+Z_T\big\rangle\\ &= \max_{q_T\in Q_T}\big\langle q_T, Z_1+Z_2+\cdots+Z_T\big\rangle,\end{aligned} \tag{27}$$

as postulated. □

6 Duality and Decomposition

An advantage of formula (22) is that its right hand side remains well-defined also for sequences $\{Z_t\}$, which are not adapted to the filtration $\{\mathscr{F}_t\}$. This allows for the development of the corresponding duality theory and decomposition.

Consider the extended problem formulation corresponding to to the risk-neutral formulation (13). The nonanticipativity constraints (10) can be compactly written as a system of linear equations $x = \Pi x$, where Π is the projection on the implementable subspace:

$$\Pi(x_1,\dots,x_T) = \big(\mathbb{E}x_1, \mathbb{E}[x_2|\mathscr{F}_2],\dots,\mathbb{E}[x_{T-1}|\mathscr{F}_{T-1}], x_T\big).$$

Employing the dual representation of the dynamic measure of risk $\rho(\cdot)$, we obtain the following problem:

$$\min_x \max_{q\in\mathcal{Q}_T} \sum_{s\in\Omega} q^s \langle c^s, x^s\rangle \tag{28}$$

$$\text{s.t. } x - \Pi x = 0, \tag{29}$$

$$x^s \in \mathscr{X}^s, \quad s\in\Omega. \tag{30}$$

We write $\langle c^s, x^s\rangle$ for the sum $\sum_{t=1}^T \langle c_t^s, x_t^s\rangle$. By Theorem 1, this problem is equivalent to the problem of minimizing (4), subject to (2) and (1).

We now develop duality relations for problem (28)–(30), extending to the risk-averse case the approach outlined in [31, Sec. 3.2.4]. After associating Lagrange multipliers λ with the nonanticipativity constraints (29), we obtain the following Lagrangian function:

$$L(x,\lambda) = \max_{q\in\mathcal{Q}_T} \sum_{s\in\Omega} \left(q^s\langle c^s, x^s\rangle + p^s\langle \lambda^s, x^s - \Pi^s x\rangle\right).$$

It is sufficient to consider λ such that $\Pi\lambda = 0$, because any shift of λ by by a vector in the range of Π does not affect the last term. More specifically, we require that

$$\sum_{s\in\mathscr{S}(\nu)} p^s \lambda_t^s = 0, \quad \nu\in\Omega_t, \quad t = 1,\dots,T-1. \tag{31}$$

Under this condition, the Lagrangian simplifies:

$$L(x,\lambda) = \max_{q\in\mathcal{Q}_T} \sum_{s\in\Omega} \left(q^s\langle c^s, x^s\rangle + p^s\langle \lambda^s, x^s\rangle\right). \tag{32}$$

The dual function is defined as follows:

$$L_D(\lambda) = \inf_{x\in\mathscr{X}} L(x,\lambda),$$

and the dual problem is to find

$$\max_{\Pi\lambda=0} \inf_{x\in\mathscr{X}} \max_{q\in\mathcal{Q}_T} \sum_{s\in\Omega} \left(q^s\langle c^s, x^s\rangle + p^s\langle \lambda^s, x^s\rangle\right). \tag{33}$$

The function under the "max – inf – max" operations is bilinear in x and q, the set Q_T is convex and compact, and the set $\mathscr{X}$ is convex. Therefore, we can interchange the inner "inf" and "max" operations to write the dual problem as follows:

$$\max_{\Pi\lambda=0} \max_{q\in Q_T} \left[\inf_{x\in\mathscr{X}} \sum_{s\in\Omega} \left(q^s \langle c^s, x^s \rangle + p^s \langle \lambda^s, x^s \rangle \right) \right]. \tag{34}$$

It is convenient to replace the measure q with its density δ with respect to p. Clearly, δ lives in a convex compact set

$$\Delta = \left\{ \delta \in \mathbb{R}^{|\Omega|} : \left(p^s \delta^s \right)_{s\in\Omega} \in Q_T \right\}. \tag{35}$$

The dual problem takes on the form:

$$\max_{\Pi\lambda=0} \max_{\delta\in\Delta} \left[\inf_{x\in\mathscr{X}} \sum_{s\in\Omega} p^s \left(\delta^s \langle c^s, x^s \rangle + \langle \lambda^s, x^s \rangle \right) \right]. \tag{36}$$

The problem in brackets has the same structure as in the risk-neutral case, but with scenario costs re-scaled by δ^s.

Theorem 2 *If Problem* (28)-(30) *has an optimal solution then the dual problem* (36) *has an optimal solution, and the optimal values of both problems coincide.*

The theorem follows from the duality theory in convex programming (see, e.g., [24, Thms. 4.7 and 4.8]). No constraint qualification is needed, because the constraints (29) are linear and the sets $\mathscr{X}^s$, $s \in \Omega$, are convex closed polyhedra.

Observe that the inner problem (in brackets) in (36) decomposes into individual scenario subproblems

$$\min_{x^s\in\mathscr{X}^s} \langle \delta^s c^s + \lambda^s, x^s \rangle, \quad s \in \Omega. \tag{37}$$

These subproblems can be readily solved by specialized techniques, exploiting the structure of the deterministic version of the dynamic problem in question.

Our approach can be interpreted as a construction of a family of risk-neutral approximations of the problem, one for each $\delta \in \Delta$.

7 Master Problem

Let us denote by $\Psi^s(\lambda^s, \delta^s)$ the optimal value of problem (37). The main difficulty is to solve the dual problem:

$$\max_{\Pi\lambda=0} \max_{\delta\in\Delta} \sum_{s\in\Omega} p^s \Psi^s(\lambda^s, \delta^s). \tag{38}$$

As each $L^s(\cdot,\cdot)$ is concave and piecewise-linear, problem (38) is a convex programming problem.

The optimal value of the scenario subproblem (37) is a composition of the linear map $(\lambda^s, \delta^s) \mapsto \delta^s c^s + \lambda^s$ with the support function of the set $\mathscr{X}_s$. Using rules of subdifferential calculus we obtain

$$\partial \Psi^s(\lambda^s, \delta^s) = \left\{ \left(x^s, \langle c^s, x^s \rangle\right) : x^s \text{ is a solution of (37)} \right\}. \tag{39}$$

As the objective of (38),

$$D(\lambda, \delta) = \sum_{s \in \Omega} p^s \Psi^s(\lambda^s, \delta^s),$$

is a sum of terms that have no variables in common, we get

$$\partial D(\lambda, \delta) = \partial \Psi^1(\lambda^1, \delta^1) \times \cdots \times \partial \Psi^{|\Omega|}(\lambda^{|\Omega|}, \delta^{|\Omega|}). \tag{40}$$

Therefore, to calculate a subgradient at a point (λ, δ) we need to solve subproblems (37) and apply formula (40). In principle, problem (38) can be solved by any nonsmooth optimization method. One simple possibility would be the cutting plane method (see, e.g., [12,24]); another choice is the bundle method (see [12,14,15,24]).

The essence of the bundle method is the application of regularization with respect to the decision variables, which are in our case λ and δ, similarly to the proximal point method. This allows to localize the iterations and makes the bundle method more reliable for problems of higher dimension, where the cutting plane method becomes very slow.

Here, the specificity of problem (38) is that regularization is mainly needed for the nonanticipativity multipliers λ. The densities δ are restricted to live in a compact set Δ; in the extreme case of the risk-neutral problem we simply have $\Delta = \{(1,1,\ldots,1)\}$. We therefore propose a *partial bundle method*, which employs regularization with respect to the variables λ only. Exactly as the bundle method, it collects for every scenario s optimal solutions x^{sj} of the scenario subproblems and corresponding solutions $(\lambda^{sj}, \delta^{sj})$ of the master problem at iterations $j \in J_s$. The set J_s may be the set of all previous iterations, or its subset determined by the cut selection rules of the bundle method. The method also has the regularization center $\bar{\lambda}$, which is updated depending on the success of the current iteration, and uses a regularization coefficient $r > 0$.

The master problem of the partial bundle method has the following form

$$\begin{aligned} \max_{v_s, \lambda, \delta} \; & \sum_{s \in \Omega} p^s \left(v_s - \frac{r}{2} \|\lambda^s - \bar{\lambda}^s\|^2 \right) \\ \text{s.t. } & v_s \le \left\langle \delta^{sj} c^s + \lambda^{sj}, x^{sj} \right\rangle + \left\langle \left(x^{sj}, \langle c^s, x^{sj} \rangle\right), (\lambda^s, \delta^s) - (\lambda^{sj}, \delta^{sj}) \right\rangle, \\ & \qquad s \in \Omega, \quad j \in J_s, \\ & \Pi \lambda = 0, \\ & \delta \in \Delta. \end{aligned} \tag{41}$$

After its solution, the regularization center $\bar{\lambda}$, the regularization coefficient r, and the sets of cuts are updated in exactly the same way as in the bundle method (see [15, 24]). Convergence analysis of the partial bundle method is nontrivial and lengthy. As these details would take us far away from the main topic of our presentation; we outline the analysis in the Appendix, for the basic problem of minimizing a convex

function of two decision vectors, without the complications of dealing with the sum of functions, over $s \in \Omega$. Our master problem (41) uses disaggregated subgradients, as in [5,22]: each v_s is an upper bound on the corresponding function $\Psi^s(\lambda^s, \delta^s)$. The interested readers are referred to [8].

8 Numerical Illustration

8.1 The Model

Our aim is to illustrate the scenario decomposition approach and the methods discussed in previous sections on the following inventory and assembly problem. A product line consists of several different models. Each model has its own list of parts, but different models may have some parts in common. At the first stage, we decide how many units of each part will be bought. After the purchase is done, the actual demand for the different models is revealed. Then we decide how many units of each model will be produced, while keeping within the constraints defined by the numbers of parts available.

There is a penalty for each unit of unsatisfied demand and there is a "storage cost" associated to each unit that is produced over the demand. The storage cost involves product depreciation and is a random variable which will become known only after the second stage decisions have been made. It is assumed that all the products will eventually be sold and the storage cost is paid only once.

Let z_i be the number of parts of type i that will be purchased and let u_j be the number of units of model j that will be produced. Let M be the integer nonnegative matrix that describes the parts needed to assemble each different model, i.e. Mu is the vector of parts necessary to assemble the vector of models u. Random demand for product j is denoted by D_j and random unit storage cost is denoted by H_j. Other problem parameters, which are deterministic, are: r_j - selling price of product j, c_i - cost of part i, l_j - penalty for uncovered demand of product j.

Our goal is to minimize the negative of the profit, which is composed of three parts: $Z_1 = \sum_i c_i z_i$, $Z_2 = -\sum_j r_j u_j$, and $Z_3 = \sum_j [l_j(D_j - u_j)_+ + H_j(u_j - D_j)_+]$. Since the components Z_2 and Z_3 are random, and our decisions u depend on the demand vector observed, we express the production problem as a three stage risk-averse optimization problem. In fact, there are no third stage decisions: only random cost evaluation.

At stages 1 and 2 we use the conditional mean–semideviation risk measures of the first order of the form (3) with coefficients $\kappa_1 \in [0,1]$ and $\kappa_2 \in [0,1]$, respectively

Assume that there are N possible demand realizations each occurring with corresponding probability p_s. Moreover, suppose that each demand realization s there are N_s possible storage cost realizations each occurring with probability $p_{s\eta}$, $\eta = 1,\ldots,N_s$. For given decisions u^s at node s, the cost equals:

$$Z_2^s + Z_3 = -\langle r, u^s \rangle + \langle l, w^s \rangle + \langle H^{s\eta}, v^s \rangle,$$

where w^s and v^s are the under and over production due to decision y^s at node s. In this case a straightforward linear programming formulation of the problem is the

following:

$$\begin{aligned}
\min_{\substack{z,u,w,v\\ \rho,\sigma,\zeta,\gamma}} \ & \langle c,z\rangle + \sum_{s=1}^{N} p_s \rho^s + \kappa_1 \sum_{s=1}^{N} p_s \sigma^s \\
\text{s.t. } & \rho^s = \sum_{\eta=1}^{N_s} p_{s\eta}\, \zeta^{s\eta} + \kappa_2 \sum_{\eta=1}^{N_s} p_{s\eta}\, \gamma^{s\eta}, \\
& \sigma^s \geq \rho^s - \sum_{k\in\Omega_2} p_k \rho^k, \quad \sigma^s \geq 0, \\
& \zeta^{s\eta} = -\langle r, u^s\rangle + \langle l, w^s\rangle + \langle H^{s\eta}, v^s\rangle, \\
& \gamma^{s\eta} \geq \zeta^{s\eta} - \sum_{k=1}^{N_s} p_{sk}\, \zeta^{sk}, \quad \gamma^{s\eta} \geq 0, \\
& Mu^s - z \leq 0, \quad u^s \geq 0, \\
& w^s \geq D^s - u^s, \quad w^s \geq 0, \\
& v^s \geq u^s - D^s, \quad v^s \geq 0, \\
& \text{for all } s = 1,\ldots,N \text{ and } \eta = 1,\ldots,N_s.
\end{aligned} \tag{42}$$

In the problem above, $D^s := (D_1^s,\ldots,D_m^s)$ is the sth realization of product demands, and $H_i^{s\eta}$ is the storage cost of product i under demand realization s and storage realization η. The variable ρ^s represents the value of the conditional risk measure $\rho_2(Z_2+Z_3)$ at node s, and the value of the risk measure $\rho_1(\cdot)$ is calculated directly in the objective function. The variables ζ represent cost realizations in the corresponding scenarios. The variables σ and γ represent the upper semideviations of the costs at stage 1 and 2, respectively.

The size of the linear programming representation of the production problem shows the importance of developing efficient methods to solve multi stage risk-averse problems. We applied to our problem the cutting plane, the classical bundle, and the partial bundle method. Whenever possible, we compared the results obtained by these methods with the result of solving the linear programming problem (42) directly by a simplex algorithm. For the scenario decomposition methods, we considered two versions. One was the full three-stage version, which is most general and applies also to problems involving decisions at the last stage and general non-polyhedral measures of risk. Another version was a model with a truncated two-stage tree, in which the problems at the second stage are risk-averse problems themselves. This was possible due to the polyhedral structure of the mean–semideviation risk measure and to the absence of third stage decisions.

8.2 The Partial Bundle Method

To obtain explicitly the master problem of the partial bundle method for our application we need to calculate the set Δ appearing in (41). The structure of the subdifferential of the mean upper semideviation is well known (see [31] page 278), namely,

$$\partial\rho_1(0) = \left\{ \mathbb{1} - \mathbb{1}\sum_{s=1}^{N} p_s\tau_s + \tau \;\middle|\; \tau = (\tau_s)_{s=1}^{N} \text{ and } 0 \leq \tau_s \leq \kappa_1 \right\} \tag{43}$$

and

$$\partial\rho_2^s(0) = \left\{ \mathbb{1} - \mathbb{1}\sum_{\eta=1}^{N_s} p_{s\eta}\iota_{s\eta} + \iota^s \;\middle|\; \iota^s = (\iota_{s\eta})_{\eta=1}^{N_s} \text{ and } 0 \le \iota_{s\eta} \le \kappa_2 \right\}, \tag{44}$$

where $\mathbb{1}$ is the vector with all entries equal to 1. Let $\partial\rho_2(0) := \partial\rho_2^1(0) \times \cdots \times \partial\rho_2^N(0)$ and $\pi = (p_{s\eta})_{s\in\Omega_1,\eta\in C(s)}$. Then $Q_2 = \mathscr{A}_2 \circ \mathscr{A}_1 = \partial\rho_2(0) \circ \partial\rho_1(0)$ and $\Delta = \left\{ \delta : \left(p^s\delta^s\right)_{s=1}^N \in Q_2 \right\}$. Thanks to the structure of the subdifferentials (43) and (44) the set Δ is polyhedral, and so, $\Delta = \left\{ \left(\delta_\eta^s\right)_{s=1,\ldots,N,\eta=1,\ldots,N_s} \right\}$ such that

$$\begin{aligned}
\delta_\eta^s &= p_{s\eta}\left[1 - \sum_{k=1}^{N} p_k\tau_k + \tau_s - \sum_{k=1}^{N_s} p_{sk}\varepsilon_{sk} + \varepsilon_{s\eta}\right] \\
&\quad 0 \le \tau_i \le \kappa_1, \quad i = 1,\ldots,N, \\
&\quad 0 \le \varepsilon_{ij} \le \kappa_2\left(1 - \sum_{k=1}^{N_s} p_k\tau_k + \tau_i\right), \quad i = 1,\ldots,N, \quad j = 1,\ldots,N_s.
\end{aligned}$$

The master problem of the partial bundle method for our application is:

$$\begin{aligned}
\max_{v_s,\lambda,\delta}\ & \sum_{s=1}^{N} p^s\left(v_s - \frac{r}{2}\|\lambda^s - \bar{\lambda}^s\|^2\right) \\
\text{s.t. } & v_s \le \left\langle \delta^{sj}c^s + \lambda^{sj}, x^{sj}\right\rangle + \left\langle \left(x^{sj}, \langle c^s, x^{sj}\rangle\right), (\lambda^s,\delta^s) - (\lambda^{sj},\delta^{sj})\right\rangle, \\
& \Pi\lambda = 0, \\
& \delta^s = \left(p_{s\eta}\left[1 - \sum_{k=1}^{N} p_k\tau_k + \tau_s - \sum_{k=1}^{N_s} p_{sk}\varepsilon_{sk} + \varepsilon_{s\eta}\right]\right)_{\eta=1}^{N_s}, \\
& 0 \le \tau_s \le \kappa_1, \\
& 0 \le \varepsilon_{s\eta} \le \kappa_2\left(1 - \sum_{k=1}^{N_s} p_k\tau_k + \tau_s\right), \quad \eta = 1,\ldots,N_s, \\
& \text{for all } j \in J_s, \quad s = 1,\ldots,N.
\end{aligned} \tag{45}$$

At every iteration j of the partial bundle method the obtained subgradient have the following form

$$\left[\left(p_1 z^{1j}\right)^\top,\ldots,\left(p_N z^{Nj}\right)^\top, p_1\langle c^1, z^{1j}\rangle,\ldots,p_N\langle c^N, z^{Nj}\rangle, \left(p_1 G^1 y^{1j}\right)^\top,\ldots,\left(p_N G^N y^{Nj}\right)^\top\right],$$

where $x^{sj} := (z^{sj}, y^{sj})$ is the optimal solution of subproblem (37) for scenario s at iteration j with z^{sj} corresponding to the first stage components of x^{sj}, and y^{sj} corresponding to the second and third stage components of x^{sj}. Also, c^s is the cost vector of the first stage scenario s, and G^s is the matrix of second stage scenario costs corresponding to the first stage scenario s. In our example $c^s = c$ and the rows of G^s are $(g_{s_\eta})^\top = \left(r^\top, l^\top, (H^{s\eta})^\top\right)$, for every $s = 1,\ldots,N$, $\eta = 1,\ldots,N_s$.

After a few algebraic simplifications we derive from (52) the individual scenario subproblems for each scenario $s = 1, \ldots, N$,

$$\begin{aligned} \min_{z,u,w,v,\zeta} \ & \alpha^s \langle c^s, z\rangle + \langle \beta^s, \zeta\rangle + \langle \lambda^s, z\rangle \\ \text{s.t. } & \zeta^\eta = -\langle r, u\rangle + \langle l, w\rangle + \langle H^{s\eta}, v\rangle, \quad \eta = 1, \ldots, N_s, \\ & Mu - z \le 0, \quad u \ge 0, \\ & w \ge D^s - u, \quad w \ge 0, \\ & v \ge u - D^s, \quad v \ge 0, \end{aligned} \tag{46}$$

where each y^{sj} component of x^{sj} in (45) has been subdivided according to (42), i.e., $x := (z, y) := (z, u, v, w)$. Similarly, α^s, β^s are the corresponding z, y components of δ^s.

8.3 The Truncated Tree Method

In order to obtain the truncated two-stage tree method we need to find an efficient way of evaluating the second stage upper semideviation risk measure. Applying (17), we obtain for every $s = 1, \ldots, N$,

$$\rho_2^s(G^s y) = \max_{\delta \in \partial \rho_2^s(0)} \sum_{\eta=1}^{N_s} \delta_\eta p_{s\eta} g_{s\eta}^\top y, \tag{47}$$

where $\partial \rho_2^s(0)$ is obtained from (44). Substituting (44) into (47) gives

$$\rho_2^s(G^s y) = \max_{\iota \in [0, \kappa_2]^{N_s}} \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} \iota_\eta p_{s\eta} \left[g_{s\eta}^\top y - \sum_{\zeta=1}^{N_s} p_{s\zeta} g_{s\zeta}^\top y \right]. \tag{48}$$

Therefore $\rho_2^s(G^s y)$ can be obtained by solving the following linear program

$$\begin{aligned} \min \ & \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} d_\eta \\ \text{s.t. } & d_\eta \ge \kappa_2 p_{s\eta} \left[g_{s\eta}^\top y - \sum_{\zeta=1}^{N_s} p_{s\zeta} g_{s\zeta}^\top y \right], \quad \eta = 1, \ldots, N_s, \\ & d_\eta \ge 0, \quad \eta = 1, \ldots, N_s. \end{aligned} \tag{49}$$

The main idea of the truncated tree method is that instead of minimizing (4) subject to (1) and (2), we minimize

$$\widetilde{\rho_{1,3}} = Z_1 + \rho_1 \left(\widetilde{Z_2} \right), \tag{50}$$

subject to (1) and (2), and

$$\widetilde{Z_2} = Z_2 + \rho_2 (Z_3). \tag{51}$$

We consider the truncated problem as a two-stage problem and apply to it the same dual analysis that we did before. At the end we obtain formulation (38) with a few

key differences. First, λ and δ refer to the random variables Z_1 and $\widetilde{Z}_2$ and have no components directly relating to either Z_2 or Z_3. More importantly, the individual scenario subproblems should take into consideration the cost of new random variable $\widetilde{Z}_2$ and thus (37) is replaced by

$$\min_{(z^s,y^s)\in\mathscr{X}^s} \langle \delta^s c^s + \lambda^s, z^s\rangle + \delta^s \rho_2^s(g_s^\top y^s), \quad s=1,\ldots,N, \tag{52}$$

where z^s and y^s are the decision variables corresponding to the first and second stage scenarios. By substituting (49) and (52) into (38) we obtain the following problem formulation for our application

$$\max_{\Pi\lambda=0} \max_{\delta\in\Delta} \sum_{s=1}^{N} p^s \Psi^s(\lambda^s,\delta^s), \tag{53}$$

where $\Psi^s(\lambda^s,\delta^s)$ is the optimal value of the following problem

$$\begin{aligned}
\min_{z,y,d}\ & \delta^s \left[(c^s)^\top z + \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} d_\eta \right] + (\lambda^s)^\top z \\
\text{s. t. } & d_\eta \geq \kappa_2 p_{s\eta} \left[g_{s\eta}^\top y - \sum_{\zeta=1}^{N_s} p_{s\zeta} g_{s\zeta}^\top y \right], \quad \eta = 1,\ldots,N_s, \\
& B_3^s z + A_3^s y = b_3^s, \\
& z\in X,\ y\geq 0,\ d\geq 0.
\end{aligned} \tag{54}$$

At every iteration j of the truncated tree partial bundle method, the subgradient has the following form

$$\left[\left(p_1 z^{1j}\right)^\top, \ldots, \left(p_N z^{Nj}\right)^\top, p_1 \phi_s\left(z^{1j}, y^{1j}, d^{1j}\right), \ldots, p_N \phi_s\left(z^{N,j}, y^{Nj}, d^{Nj}\right) \right], \tag{55}$$

where for every $s=1,\ldots,N$,

$$\phi_s(z,y,d) = (c^s)^\top z + \sum_{\eta=1}^{N_s} p_{s\eta} g_{s\eta}^\top y + \sum_{\eta=1}^{N_s} d_\eta,$$

and $x^{sj} := (z^{sj}, y^{sj}, d^{sj})$ is the optimal solution of subproblem (54) for scenario s at iteration j.

By construction, we only consider the first scenarios for the decomposition in (53) and so $\Delta = \left\{ \delta : \left(p^s\delta^s\right)_{s=1}^N \in \partial\rho_1(0) \right\}$, where $\partial\rho_1(0)$ was shown in (43). Therefore the master problem of partial bundle method for the truncated tree method has the

following form:

$$\begin{aligned}
&\max_{v,\lambda,\delta} \sum_{s=1}^{N} p^s \left(v_s - \frac{r}{2}\|\lambda^s - \bar{\lambda}^s\|^2 \right) \\
&\text{s.t. } v_s \le \left\langle \delta^{sj} c^s + \lambda^{sj}, x^{sj} \right\rangle + \left\langle \left(x^{sj}, \langle c^s, x^{sj}\rangle\right), (\lambda^s, \delta^s) - (\lambda^{sj}, \delta^{sj}) \right\rangle, \\
&\quad \Pi\lambda = 0, \\
&\quad \delta^s = p_s \left[1 - \sum_{k=1}^{N} p_k \tau_k + \tau_s \right], \\
&\quad 0 \le \tau_s \le \kappa_1, \\
&\quad \text{for all } j \in J_s, \quad s = 1, \ldots, N.
\end{aligned} \tag{56}$$

After a few algebraic simplifications we derive from (52) the individual truncated tree scenario subproblems for each scenario $s = 1, \ldots, N$,

$$\begin{aligned}
&\min_{z,u,w,v,\zeta} \delta^s t + \langle \lambda^s, z \rangle \\
&\quad \text{s.t. } \zeta^\eta = -\langle r, u \rangle + \langle l, w \rangle + \langle H^{s\eta}, v \rangle, \\
&\qquad t = \langle c^s, z \rangle + \left\langle (p_{sk})_{k=1}^{N_s},\ \zeta + \kappa_2 S \right\rangle \\
&\qquad S_\eta \ge \zeta^\eta - \left\langle (p_{sk})_{k=1}^{N_s},\ \zeta \right\rangle, \quad S_\eta \ge 0, \\
&\qquad Mu - z \le 0, \quad u \ge 0, \\
&\qquad w \ge D^s - u, \quad w \ge 0, \\
&\qquad v \ge u - D^s, \quad v \ge 0, \\
&\qquad \text{for all } \eta = 1, \ldots, N_s,
\end{aligned} \tag{57}$$

where each decision variable y^{sj} from (52) has been subdivided according to (42), i.e. $x := (z, y) := (z, u, v, w)$.

Following this we compared the total running time, total number of iterations, and the average time per iteration of each method. Table 1 shows the comparison of all the methods on a problem with 10 parts and 5 products, for different numbers of first-stage and second-stage scenarios.

The classical cutting plane method was inefficient and failed to converge in a reasonable time on most instances, while being outperformed by all the other methods when it converged. For this reason we omitted it from Table 1. Clearly, small problems are best solved directly by linear programming in formulation (42). The usefulness of decomposition is shown when we consider large problems. For example on the instance with 200 first-stage scenarios, with each followed by 200 second-stage scenarios, the general bundle and the partial truncated tree methods outperformed the linear programming formulation. More important is the case with 300 first-stage scenarios, with 300 second-stage scenarios after each of them, where the linear programming approach failed, but the truncated tree and partial truncated tree methods

Size	LP	Truncated Tree			Partial Truncated Tree			General Bundle		
$N \times N_s$	Time	Time	Iter.	T/I	Time	Iter.	T/I	Time	Iter.	T/I
6×3	0	106	476	0.223	15	97	0.155	84	492	0.171
5×5	0	95	451	0.211	36	194	0.186	61	419	0.146
5×6	0	75	388	0.193	13	86	0.151	48	270	0.178
6×6	0	134	574	0.233	133	441	0.302	109	521	0.209
10×10	0	313	435	0.720	287	419	0.685	309	501	0.617
50×50	5	1381	510	2.708	1652	485	3.406	3283	414	7.930
100×100	98	5570	660	8.439	1547	300	5.157	28316	579	48.91
200×200	5767	5975	240	24.89	4722	200	23.61	54336	291	186.7
300×300	-	19910	255	78.08	20622	255	80.87	-	-	-

Table 1 Performance of decomposition methods. Tests were performed for N first-stage scenarios, and N_s second-stage scenarios following each first-stage scenario.

were able to find a solution. In this case the meager memory requirements of these methods allowed us to obtain a solution even when the linear programming formulation was too large for our computer memory. In general, we saw the partial truncated tree method outperforming the truncated tree method but this might be problem specific.

Notice that the truncated tree method moves the calculation of the second stage risk measure from the master problem to the subproblems resulting in a smaller master problem but larger subproblems. This is the main difference between the truncated tree and general bundle methods. In larger instances, the dimension of the master problem affects the number of iterations necessary to find a solution, as well as time to solve the master problem at each iteration. For these reason, the truncated tree method with its simpler master problem outperforms the general bundle method on the largest instances.

Acknowledgements This research was supported by the NSF award CMII-0965689.

References

1. P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, *Coherent measures of risk*, Math. Finance 9 (1999) 203–228.
2. P. Artzner, F. Delbaen, J.-M. Eber, D. Heath, and H. Ku, *Coherent multiperiod risk adjusted values and Bellmans principle*, Annals of Operations Research 152 (2007) 5–22.
3. P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath,, *Coherent measures of risk*, Risk management: value at risk and beyond (Cambridge, 1998), Cambridge Univ. Press, Cambridge, 2002, pp. 145–175.
4. J.-P. Aubin and H. Frankowska, *Set-Valued Analysis*, Birkhäuser, Boston, 1990.
5. J. R. Birge and F. V. Louveaux, *A multicut algorithm for two-stage stochastic linear programs*, European Journal of Operational Research, 34 (1988) 384–392.
6. Birge, J. R., and F. V. Louveaux, *Introduction to Stochastic Programming*, Springer, New York, 1997.
7. P. Cheridito, F. Delbaen, and M. Kupper, *Dynamic monetary risk measures for bounded discrete-time processes*, Electronic Journal of Probability 11 (2006) 57–106.
8. R. A. Collado, *Scenario Decomposition of Risk-Averse Mustistage Stochastic Programming Problems*, PhD Dissertation, Rutgers University, 2010.
9. H. Föllmer and A. Schied, *Stochastic Finance. An Introduction in Discrete Time*, Walter de Gruyter, Berlin, 2004.

10. M. Fritelli and E. Rosazza Gianin, *Putting order in risk measures*, Journal of Banking and Finance 26 (2002) 1473–1486.
11. M. Fritelli and G. Scandolo, *Risk measures and capital requirements for processes*, Mathematical Finance 16 (2006) 589–612.
12. J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms*, Springer, Berlin, 1993.
13. P. Kall and J. Mayer, *Stochastic Linear Programming*, Springer, New York, 2005.
14. K. C. Kiwiel, *An aggregate subgradient method for nonsmooth convex minimization*, Mathematical Programming 27 (1983) 320–341.
15. K. C. Kiwiel, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics, vol. 1133, Springer-Verlag, Berlin, 1985.
16. N. Miller and A. Ruszczyński, *Risk-averse two-stage stochastic linear programming: modeling and decomposition*, Operations Research, to appear in 2011.
17. J.M. Mulvey and A. Ruszczyński, *A new scenario decomposition method for large-scale stochastic optimization*, Operations Research 43 (1995) 477–490.
18. G. Ch. Pflug and W. Römisch. *Modeling, Measuring and Managing Risk*. World Scientific, Singapore, 2007.
19. A. Prékopa, *Stochastic Programming*, Kluwer, Dordrecht, 1995.
20. F. Riedel, *Dynamic coherent risk measures*, Stochastic Processes and Their Applications, 112 (2004) 185–200.
21. R. T. Rockafellar, S. Uryasev and M. Zabarankin, *Deviation measures in risk analysis and optimization*, Finance and Stochastics 10 (2006) 51–74.
22. A. Ruszczyński, *A regularized decomposition method for minimizing a sum of polyhedral functions*, Mathematical Programming 35 (1986) 309–333.
23. A. Ruszczyński, *Decomposition methods*, in: Stochastic Programming, Handbooks Oper. Res. Management Sci., vol. 10, Elsevier, Amsterdam, 2003, pp. 141-211.
24. A. Ruszczyński, *Nonlinear Optimization*, Princeton University Press, Princeton, NJ, USA, 2006.
25. A. Ruszczyński, Risk-averse dynamic programming for Markov decision processes, *Mathematical Programming*, to appear in 2010.
26. A. Ruszczyński and A. Shapiro, *Optimization of risk measures*, In: *Probabilistic and Randomized Methods for Design under Uncertainty*, G. Calafiore and F. Dabbene (Eds.), Springer, London, 2005.
27. A. Ruszczyński and A. Shapiro, *Optimization of convex risk functions*, Math. Oper. Res. 31 (2006) 433–452.
28. A. Ruszczyński and A. Shapiro, *Conditional risk mappings*, Math. Oper. Res. 31 (2006) 544–561.
29. A. Ruszczyński and A. Shapiro, *Corrigendum to: "Optimization of convex risk functions," mathematics of operations research31 (2006) 433–452*, Math. Oper. Res. 32 (2007) 496–496.
30. G. Scandolo, *Risk Measures in a Dynamic Setting*, PhD Thesis, Università degli Studi di Milano, Milan, 2003.
31. A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*, MPS-SIAM Series on Optimization, No. 9, MPS-SIAM, Philadelphia, 2009.

A The Partial Bundle Method

A.1 The Method

We consider the problem

$$\underset{(x,y)\in A}{\text{minimize}}\ f(x,y), \tag{58}$$

in which the set $A \subseteq \mathbb{R}^n \times \mathbb{R}^m$ is closed convex and the function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex, proper, and lower semicontinuous. We assume that the set A is x-bounded in the sense that for every bounded subset $Y \subset \mathbb{R}^m$ the intersection $A \cap (\mathbb{R}^n \times Y)$ is bounded.

We define the following *regularized master problem*:

$$\underset{(x,y)\in A}{\text{minimize}}\ \frac{r}{2}\|y - w^k\|^2 + f^k(x,y), \tag{59}$$

where the model f^k is defined by:

$$f^k(x,y) \triangleq \max_{j\in J_k}\left[f(x^j,y^j)+\left\langle g^j,(x,y)-(x^j,y^j)\right\rangle\right], \tag{60}$$

with $g^j \in \partial f(x^j,y^j)$, $j \in J_k$. The set J_k is a subset of $\{1,\ldots,k\}$ determined by a procedure for selecting cuts. At this moment we may think of J_k as being equal to $\{1,\ldots,k\}$.

In the proximal term $(r/2)\|y-w^k\|^2$, where $r>0$, the center (v^k,w^k) is updated depending on the relations between the value of $f(x^{k+1},y^{k+1})$ at the mater's solution, (x^{k+1},y^{k+1}), and its prediction provided be the current model, $f^k(x^{k+1},y^{k+1})$. If these values are equal or close, we set $(v^{k+1},w^{k+1}) := (x^{k+1},y^{k+1})$ (*descent step*); otherwise $(v^{k+1},w^{k+1}) := (v^k,w^k)$ (*null step*). In any case, the collection of cuts is updated, and the iteration continues.

The regularized master problem can be equivalently written as a problem with a quadratic objective function and linear constraints:

$$\begin{aligned} &\text{minimize } \; z+\frac{r}{2}\|y-w^k\|^2 \\ &\text{subject to } \; z \geq f(x^j,y^j)+\left\langle g^j,(x,y)-(x^j,y^j)\right\rangle, \; j\in J_k, \\ &\qquad\qquad (x,y)\in A. \end{aligned} \tag{61}$$

If the set A is a convex polyhedron, the master problem can be readily solved by specialized techniques, enjoying the finite termination property.

Let us observe that problem (61) satisfies Slater's constraint qualification condition. Indeed, for every $(x_S,y_S)\in A$ we can choose z_s so large that all constraints are satisfied as strict inequalities. Therefore the optimal solution of the master problem satisfies the necessary and sufficient conditions of optimality with Lagrange multipliers (see [24, Thm. 3.34]). We denote by $\lambda_j^k \in J_k$, the Lagrange multipliers associated with the constraints of problem (61).

The detailed algorithm is stated below. The parameter $\gamma\in(0,1)$ is a fixed constant used to compare the observed improvement in the objective value to the predicted improvement.

Step 0. Set $k:=1$, $J_0:=\emptyset$, and $z^1:=-\infty$.

Step 1. Calculate $f(x^k,y^k)$ and $g^k\in\partial f(x^k,y^k)$. If $f(x^k,y^k)>z^k$ then set $J_k:=J_{k-1}\cup\{k\}$; otherwise set $J_k:=J_{k-1}$.

Step 2. If $k=1$ or if

$$f(x^k,y^k)\leq(1-\gamma)f(v^{k-1},w^{k-1})+\gamma f^{k-1}(x^k,y^k),$$

then set $(v^k,w^k):=(x^k,y^k)$; otherwise Step 2 is a null step and we set $(v^k,w^k):=(v^{k-1},w^{k-1})$.

Step 3. Solve the master problem (61). Denote by (x^{k+1},y^{k+1}) and z^{k+1} its solution and set $f^k(x^{k+1},y^{k+1}):=z^{k+1}$.

Step 4. If $f^k(x^{k+1},y^{k+1})=f(v^k,w^k)$ then stop (the point (v^k,w^k) is an optimal solution); otherwise continue.

Step 5. If Step 2 was a null step then go to Step 6. Else (Step 2 was a descent step) remove from the set of cuts J_k some (or all) cuts whose Lagrange multipliers λ_j^k at the solution of (61) are 0.

Step 6. Increase k by one, and go to Step 1.

A.2 Convergence

First we prove that if the algorithm gets stuck at a w-center then it will approximate an optimal solution.

Lemma 2 *Let f^* be an optimal solution to* (58) *and suppose that the sequence, $\{(x^k,y^k)\}$, obtained by the partial bundle method consists of only null steps from iteration t on. Then*

$$\lim_{k\to\infty} f^{k-1}(x^k,y^k)=f^*=\lim_{k\to\infty} f(x^k,y^k).$$

Proof For any $\varepsilon > 0$, let $\mathscr{K}_\varepsilon := \left\{k : k > t \text{ and } f^{k-1}(x^k,y^k) + \varepsilon < f(x^k,y^k)\right\}$ and let $k_1,k_2 \in \mathscr{K}_\varepsilon$ with $t < k_1 < k_2$.

Since we only have null steps we get that for every $k > t$, $(v^k,w^k) = (x^t,y^t)$ and the cutting plane generated at k will remain on the master problem from k on. This implies that the sequence $\left\{f^{k-1}(x^k,y^k)\right\}$ is non-decreasing from $t+1$ on. Also, since the cutting plane generated at (x^{k_1},y^{k_1}) will remain in the master problem at iteration $k_2 - 1$, we get:

$$f(x^{k_1},y^{k_1}) + \left\langle g^{k_1}, (x^{k_2},y^{k_2}) - (x^{k_1},y^{k_1})\right\rangle \leq f^{k_2-1}(x^{k_2},y^{k_2}).$$

On the other hand, $\varepsilon < f(x^{k_2},y^{k_2}) - f^{k_2-1}(x^{k_2},y^{k_2})$ which combined with the last inequality yields

$$\varepsilon < f(x^{k_2},y^{k_2}) - f(x^{k_1},y^{k_1}) + \left\langle g^{k_1}, (x^{k_1},y^{k_1}) - (x^{k_2},y^{k_2})\right\rangle.$$

Since all the steps made are null, the points y^k, with $k > t$, are contained in a bounded neighborhood of $w^k = y^t$. This and the x-boundedness of f guarantee us that $B := \operatorname{Conv}\left\{(x^j,y^j) \;\middle|\; j \in \mathscr{K}_\varepsilon\right\}$ is bounded. The function f is subdifferentiable in $\overline{B}$, so there exists a constant C such that $f(x_1,y_1) - f(x_2,y_2) \leq C\|(x_1,y_1) - (x_2,y_2)\|$, for all $x_1,x_2 \in \overline{B}$. Subgradients on bounded sets are bounded, and thus we can choose C large enough so that $\|g^j\| \leq C$, for all $j \in \mathscr{K}_\varepsilon$. It follows from the last displayed inequality that

$$\varepsilon < 2C\|(x^{k_1},y^{k_1}) - (x^{k_2},y^{k_2})\| \text{ for all } k_1,k_2 \in \mathscr{K}_\varepsilon,\ k_1 \neq k_2.$$

As the set $\overline{B}$ is compact, there can exist only finitely many points in $\mathscr{K}_\varepsilon \subset \overline{B}$ having distance at least $\varepsilon/(2C)$ from each other. Thus the last inequality implies that the set $\mathscr{K}_\varepsilon$ is finite for each $\varepsilon > 0$. This means that

$$\lim_{k\to\infty} f(x^k) - f^{k-1}(x^k) = 0. \tag{62}$$

By construction the sequences $\left\{f^{k-1}(x^k)\right\}$ and $\left\{f(x^k)\right\}$ satisfy the relation

$$f^{k-1}(x^k) \leq f^* \leq f(x^k), \text{ for every } k \in \mathbb{N}.$$

Therefore the eventual monotonicity of $\left\{f^{k-1}(x^k)\right\}$ and (62) imply that

$$\lim_{k\to\infty} f^{k-1}(x^k,y^k) = f^* = \lim_{k\to\infty} f(x^k,y^k).$$

□

Next we prove another intermediate step towards convergence.

Lemma 3 *Assume that problem* (58) *has an optimal solution and suppose that the sequence,* $\left\{(x^k,y^k)\right\}$*, obtained by the partial bundle method has infinitely many descent steps. Then the following holds.*

1. *The sequence* $\{(v^k,w^k)\}$ *approximates an optimal solution of* (58)*.*
2. *The sequence* $\left\{w^k\right\}$ *converges to a point* $\tilde{y}$ *such that there is an optimal solution of* (58) *of the form* $(\tilde{x},\tilde{y})$*.*

Proof Let us denote by $\mathscr{K}$ the set of iterations at which descent steps occur. If $(v^{k+1},w^{k+1}) = (x^{k+1},y^{k+1})$ is the optimal solution of the master problem (59), we have the necessary condition of optimality

$$0 \in \partial\left[\frac{r}{2}\|y - w^k\|^2 + f^k(x,y)\right] + N_A(x,y) \text{ at } (x,y) = (v^{k+1},w^{k+1}).$$

Hence

$$-\left[0, r(w^{k+1} - w^k)\right] \in \partial f^k(v^{k+1},w^{k+1}) + N_A(v^{k+1},w^{k+1}).$$

Let (x^*,y^*) be an optimal solution of (58). By the subgradient inequality for f^k we get (for some $h \in N_A(v^{k+1},w^{k+1})$) the estimate

$$\begin{aligned} f^k(x^*,y^*) &\geq f^k(v^{k+1},w^{k+1}) - \left\langle \left[0, r\left(w^{k+1} - w^k\right)\right], (x^*,y^*) - (v^{k+1},w^{k+1})\right\rangle \\ &\quad - \left\langle h, (x^*,y^*) - (v^{k+1},w^{k+1})\right\rangle \\ &\geq f^k(v^{k+1},w^{k+1}) - r\left\langle w^{k+1} - w^k, y^* - w^{k+1}\right\rangle. \end{aligned} \tag{63}$$

A descent step from (v^k, w^k) to (v^{k+1}, w^{k+1}) occurs, so the test of Step 2 is satisfied (for $k+1$):

$$f(v^{k+1}, w^{k+1}) \le (1-\gamma) f(v^k, w^k) + \gamma f^k(v^{k+1}, w^{k+1}).$$

After elementary manipulations we can rewrite it as

$$f^k(v^{k+1}, w^{k+1}) \ge f(v^{k+1}, w^{k+1}) - \frac{1-\gamma}{\gamma}\left[f(v^k, w^k) - f(v^{k+1}, w^{k+1})\right]. \tag{64}$$

Combining the last inequality with (63) and using the relation $f(x^*, y^*) \ge f^k(x^*, y^*)$ we obtain

$$f(x^*, y^*) \ge f(v^{k+1}, w^{k+1}) + \frac{1-\gamma}{\gamma}\left[f(v^{k+1}, w^{k+1}) - f(v^k, w^k)\right] - r\left\langle w^{k+1} - w^k, y^* - w^{k+1}\right\rangle.$$

This can be substituted to the identity:

$$\|w^{k+1} - y^*\|^2 = \|w^k - y^*\|^2 + 2\left\langle w^{k+1} - w^k, w^{k+1} - y^*\right\rangle - \|w^{k+1} - w^k\|^2.$$

After skipping the last term we get

$$\begin{aligned} \|w^{k+1} - y^*\|^2 \le\ & \|w^k - y^*\|^2 - \frac{r}{2}\left[f(v^{k+1}, w^{k+1}) - f(x^*, y^*)\right] \\ & + \frac{2(1-\gamma)}{\gamma r}\left[f(v^k, w^k) - f(v^{k+1}, w^{k+1})\right] \quad \text{for all } k \in \mathscr{K}. \end{aligned} \tag{65}$$

The series $\sum_{k=1}^{\infty}[f(v^k, w^k) - f(v^{k+1}, w^{k+1})]$ is convergent, because $\{f(v^k, w^k)\}$ is nonincreasing and bounded from below by $f(x^*, y^*)$. Therefore we obtain from (65) that the distance $\|w^{k+1} - y^*\|$ is uniformly bounded, and so $\{w^k\}$ must have accumulation points. This and the x-boundedness of f imply that the sequence $\{v^k, w^k\}$ has accumulation points.

Summing up (65) for $k \in \mathscr{K}$ we get

$$\sum_{k \in \mathscr{K}} \left(f(v^{k+1}, w^{k+1}) - f(x^*, y^*)\right) \le \frac{r}{2}\|w^1 - y^*\|^2 + \frac{1-\gamma}{\gamma}\left[f(v^1, w^1) - \lim_{k\to\infty} f(v^k, w^k)\right],$$

so $f(v^{k+1}, w^{k+1}) \to f(x^*, y^*)$, $k \in \mathscr{K}$. Consequently, at every accumulation point $(\tilde{x}, \tilde{y})$ of $\{(v^k, w^k)\}$ one has $f(\tilde{x}, \tilde{y}) = f(x^*, y^*)$. Since $(\tilde{x}, \tilde{y})$ is optimal, we can substitute it for (x^*, y^*) in (65). Skipping the negative term we get

$$\|w^{k+1} - \tilde{y}\|^2 \le \|w^k - \tilde{y}\|^2 + \frac{2(1-\gamma)}{\gamma r}\left[f(v^k, w^k) - f(v^{k+1}, w^{k+1})\right].$$

It is true not only for $k \in \mathscr{K}$ but for all k, because at $k \notin \mathscr{K}$ we have a trivial equality here. Summing these inequalities from $k = l$ to $k = q > l$ we get

$$\|w^{q+1} - \tilde{y}\|^2 \le \|w^l - \tilde{y}\|^2 + \frac{2(1-\gamma)}{\gamma r}\left[f(v^l, w^l) - f(v^{q+1}, w^{q+1})\right].$$

Since $\tilde{y}$ is an accumulation point, for $\varepsilon > 0$ we can find l such that $\|w^l - \tilde{y}\| \le \varepsilon$. Also, if l is large enough, $f(v^l, w^l) - f(v^{q+1}, w^{q+1}) \le \varepsilon$ for all $q > l$, because $\{f(v^k, w^k)\}$ is convergent. Then $\|w^{q+1} - \tilde{y}\|^2 \le \varepsilon^2 + 2\varepsilon(1-\gamma)/(\gamma r)$ for all $q > l$, so the sequence $\{w^k\}$ is convergent to $\tilde{y}$. □

Now we are ready to prove convergence of the partial bundle method.

Theorem 3 *Assume that problem* (58) *has an optimal solution, f^*, and let $\{(x^k, y^k)\}$ be the sequence obtained by the partial bundle method. Then*

$$\liminf_{k\to\infty} f(x^k, y^k) = f^*.$$

Proof If there are only finitely many descent steps then Lemma 2 gives the desired result. Thus we assume that the number of descent steps is infinite and by Lemma 3, $\lim_{k\to\infty} f(v^k, w^k) = f^*$. Clearly, the sequence $\{f(v^k, w^k)\}$ is an infinite subsequence of $\{f(x^k, y^k)\}$. Then, since $f(x^k, y^k) \ge f^*$ for every k, we obtain that $\liminf_{k\to\infty} f(x^k, y^k) = f^*$. □