A new framework to generate Lagrangian cuts in multistage stochastic mixed-integer programming

Christian Füllner[1*], X. Andy Sun[2], and Steffen Rebennack[1]

[1]Institute for Operations Research (IOR), Stochastic Optimization (SOP), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany
[*]Address correspondence to: christian.fuellner@kit.edu
[2]Sloan School of Management, Massachusetts Institute of Technology, USA

Abstract

Based on recent advances in Benders decomposition and two-stage stochastic integer programming we present a new generalized framework to generate Lagrangian cuts in multistage stochastic mixed-integer linear programming (MS-MILP). This framework can be incorporated into decomposition methods for MS-MILPs, such as the stochastic dual dynamic integer programming (SDDiP) algorithm. We show how different normalization techniques can be applied in order to generate cuts satisfying specific properties with respect to the convex hull of the epigraph of the value functions, e.g. having a maximum depth or being facet-defining. We provide computational results to evaluate the efficacy and performance of different normalizations in our new framework and compare them with existing techniques from the literature.

1 Introduction

In this paper, we study cut generation strategies that can be applied in decomposition methods for solving multi-stage stochastic mixed-integer linear programs (MS-MILP). More precisely, we present an alternative framework for the generation of Lagrangian cuts as they are used for instance in stochastic dual dynamic integer programming (SDDiP) proposed by [31].

1.1 Motivation and Prior Work

Multistage stochastic programs are very relevant to model decision-making processes in practice because often sequential decisions have to be made over a finite number of stages and under uncertainty considering the problem data of the following stages. For a large number of considered scenarios, solving these problems with standard solvers is computationally intractable, as their size grows exponentially in the number of stages. For this reason, decomposition methods exploiting their sequential and block-diagonal structure have been established as predominant solution techniques, among the most prominent ones nested Benders decomposition (BD) [6] and stochastic dual dynamic programming (SDDP) [24]. These methods decompose the large-scale problem into

stage- and scenario-specific subproblems, coupled by state variables and so-called *value functions*, denoted $Q_n(\cdot)$. For multistage stochastic linear programming problems (MS-LPs) these functions are convex polyhedral and can be exactly represented by finitely many affine functions called *cuts* [7]. However, in many applications some of the decision variables have to be integer or binary to model more complicated constraints. In this case, we obtain an MS-MILP and the value functions are in general non-convex and discontinuous.

A key challenge is that in this case, linear under-estimators are in general not tight. They may at best yield the closed convex envelope $\overline{\text{co}}(Q_n)(\cdot)$ of $Q_n(\cdot)$, which is the pointwise supremum of all affine functions majorized by $Q_n(\cdot)$ [4]. Even this property is not achieved by classical Benders cuts in general. Therefore, more focus has been put on Lagrangian cuts lately, which are constructed by solving special Lagrangian dual problems, as introduced in the original SDDiP paper [31]. These Lagrangian cuts have useful properties: They are valid under-estimators of $Q_n(\cdot)$ and they can be used to recover $\overline{\text{co}}(Q_n)(\cdot)$. As shown in [31], if all state variables of the MS-MILP are binary, this even ensures tightness for $Q_n(\cdot)$, which is sufficient to establish almost sure finite convergence of SDDiP.

Nonetheless, applying Lagrangian cuts computationally in practice comes with some considerable challenges: First, Lagrangian dual problems are often degenerate with multiple optimal solutions. Even if all the cuts associated with these solutions are tight, their approximation quality may differ significantly, as illustrated in Fig. 1. This issue is especially common for the binary state space required in SDDiP because all cuts are constructed at extreme points of the state space.

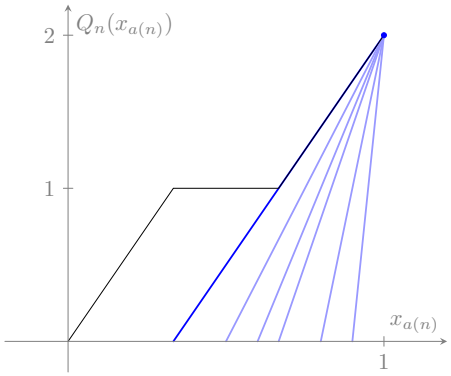


Figure 1: Tight Lagrangian cuts with different approximation quality.

Second, even if the Lagrangian dual is a convex optimization problem, it may be very costly to solve repeatedly due to its non-smooth objective function. This is only aggravated if the state space is artificially increased by a binary expansion, as it is proposed in SDDiP to ensure cut tightness in the case of non-binary state variables [31]. This computational drawback of Lagrangian cuts is already identified in the original SDDiP work [31]. It is concluded that performance-wise the improvement in cut quality is often not worth the significant increase in solution time.

Finally, tight Lagrangian cuts are crucial to ensure theoretical convergence of SDDiP, but this convergence may be quite slow. In the worst case, a complete enumeration of the binary state space is required. It is conceivable that there exist alternative, possibly non-tight, cuts that may significantly speed-up the convergence process.

For the aforementioned reasons, in this paper we address the question of how to

improve and accelerate the generation and usage of Lagrangian cuts in decomposition methods for MS-MILP such as SDDiP.

Some first attempts with this aim have been made recently. In the original SDDiP paper it is proposed to combine tight cuts with strengthened Benders cuts that are not tight in general, but outperform classical Benders cuts and are efficient to compute [31]. Rahmaniani et al. [25] present a heuristic to generate Lagrangian cuts more efficiently using inner approximations or partial relaxations. Chen and Luedtke [11] suggest to restrict the feasible set of dual multipliers to the span of Benders cut coefficients of previous iterations (without convergence guarantees for the multi-stage case).

In addition, there has been a lot of research on alternative cut generation techniques for BD, which may be applicable to the stochastic and Lagrangian setting as well. To address degeneracy and dominated cuts in BD, in their seminal paper [22] present a two-step approach to generate Pareto-optimal cuts. Their ideas are improved by [23] who shows that it is sufficient to solve a single optimization problem to generate Pareto-optimal cuts. Sherali and Lunday [27] propose to generate certain maximal non-dominated Benders cuts by solving a perturbation of the original subproblem.

A novel framework to generate Benders cuts is introduced by [17]. Its starting point is the observation that in classical BD there exists an unfavorable bias towards feasibility cuts over optimality cuts. As an alternative, the proposed framework allows to generate optimality and feasibility cuts in a unified way using the same cut generation problem. Additionally, based on this unified framework, several different cut generation techniques can be explored. More precisely, applying the framework initially leads to an unbounded separation problem. For the actual cut generation, unbounded rays have to be identified, which allows for a lot of methodological flexibility.

Fischetti et al. [17] show that using a special normalization of the cut generation problem, the obtained cuts can be proven to correspond to minimal infeasible subsystems. Hosseini and Turner [21] use a different normalization to generate *deep* cuts in BD, which are characterized by their property to maximize the distance between the separating hyperplane and the point to separate. They report considerable performance gains compared to classical BD. The idea of deep cuts is not new, but priorly discussed in context of disjunctive programming [10, 13]. Brandenberg and Stursberg [9] show how facet-defining and Pareto-optimal cuts can be generated in BD using the unified framework and the so-called reverse polar set [see also 28]. It is shown that the performance improvement using this approach is significant. The same cut generation procedure is put forward by [26] in their recent paper, but motivated from a different angle, that is geometrically. A different approach to separate facet-defining cuts is presented by [12] for disjunctive programming. To our knowledge, only the work by Fischetti et al. has been applied to the stochastic setting and to Lagrangian cuts so far [11]. The authors use the alternative framework and a specific normalization technique to generate Lagrangian cuts for two-stage stochastic MILPs.

In this paper, we provide a more general framework for the multistage case and compare various different normalization techniques from a theoretical and computational perspective. In particular, we analyze the Lagrangian cuts that are obtained if the Lagrangian dual is normalized using norm constraints or linear constraints. We show that under some assumptions, these cuts are deep, facet-defining or Pareto-optimal.

1.2 Contribution

The key contributions of this paper are summarized below.

(1) We show how the alternative cut generation framework proposed by [17] for BD can be applied to the generation of Lagrangian cuts for MS-MILPs. This idea has already been used by [11] in two-stage stochastic MILPs, but to our knowledge has not been extended to the multistage case yet.

(2) As the Lagrangian dual problems in this framework are unbounded, some normalization is required to select cut coefficients in a reasonable way. We draw on recent concepts for cut selection in BD, such as optimizing over the reverse polar set [9] or generating deep cuts [21], and extend them to the stochastic and Lagrangian setting. This way, we obtain a variety of different normalization techniques and by that generalize the cut generation approach from [11]. We show that depending on the chosen normalization, cuts satisfying different quality criteria can be obtained, *e.g.*, deep cuts, facet-defining cuts or Pareto-optimal cuts. Moreover, we investigate in detail the geometrical ideas and relations behind these normalizations.

(3) We show that linear normalizations are closely related to the identification of core points in the epigraphs $\text{epi}(Q_n)$, which can be challenging for multistage stochastic problems. Therefore, we propose five heuristic approaches for the computation of core point candidates, and thus setting up linear normalizations for the generation of Lagrangian cuts.

(4) The proposed framework for Lagrangian cut generation can be incorporated into NBD or SDDiP. We prove that under some assumptions, still (almost sure) finite convergence of these methods is guaranteed.

(5) We perform extensive computational tests for SDDiP incorporating the new cut generation framework on a capacitated lot-sizing problem from the literature. We show that the obtained lower bounds in SDDiP are majorly improved using cuts from our proposed generation framework compared to classical Lagrangian cuts or Benders cuts. We also observe that this does not necessarily guarantee an improvement of the in-sample performance of the obtained policies, though.

1.3 Structure

This paper is structured as follows. In Sect. 2 we introduce MS-MILPs formally together with our notation. In Sect. 3 we introduce the new cut generation framework for Lagrangian cuts in general. We then present different types of Lagrangian cuts that can be obtained by using special normalizations. In Sect. 4 we discuss convergence of NBD and SDDiP if these cuts are incorporated. After that, in Sect. 5, we present computational experiments for SDDiP with these new Lagrangian cuts for a capacitated lot-sizing problem from the literature. We finish with a conclusion in Sect. 6. For reasons of space, some technical proofs are shifted to Appendix B.

2 Problem Formulation

We start by introducing MS-MILPs and their decomposition formally, mostly following the notation from [31]. We consider MS-MILPS with a finite number $T \in \mathbb{N}$ of stages, where some of the problem data is uncertain and evolves according to a known stochastic process $\xi := (\xi_1, \dots, \xi_T)$ with deterministic ξ_1. We assume that the random data vectors

$\xi_t, t = 1, \dots, T$, are discrete and finite, such that the uncertainty can be modeled by a finite scenario tree. Let $\mathcal{N}$ denote the set of nodes of this tree. For each node $n \in \mathcal{N}$, the unique ancestor node is denoted by $a(n)$ and the set of child nodes is denoted by $\mathcal{C}(n)$. The probability for some node n is $p_n > 0$ and assumed to be known. The transition probabilities between adjacent nodes $n, m \in \mathcal{N}$ can then be determined as $p_{nm} := \frac{p_m}{p_n}$. For the root node r, we assume $a(r) = \emptyset$ and $p_r = 1$. We define $\overline{\mathcal{N}} := \mathcal{N} \setminus \{r\}$ to address the set of nodes without the root node, $\widetilde{\mathcal{N}}$ to address the set of nodes without leaf nodes and denote by $\mathcal{N}(t)$ the nodes at stage t.

For each node $n \in \mathcal{N}$, we distinguish state variables $x_n \in \mathbb{R}^{d_n}$, which also appear in child nodes of n, and local variables $y_n \in \mathbb{R}^{\tilde{d}_n}$. $f_n(\cdot)$ denotes the objective function of node n and $\mathcal{F}_n(x_{a(n)})$ denotes the feasible set of node n, which depends on the state variable $x_{a(n)}$ from the ancestor node. We assume that $f_n(\cdot)$ is a linear function in x_n and y_n, and that $\mathcal{F}_n(\cdot)$ is a mixed-integer polyhedral set for all $x_{a(n)}$. More precisely, we assume it to be defined by

$$\begin{aligned}\mathcal{F}_n(x_{a(n)}) := \Big\{(x_n, y_n) \in \mathbb{R}^{d_n} \times \mathbb{R}^{\tilde{d}_n} \ : x_n \in X_n,\ y_n \in Y_n, \\ A_n x_{a(n)} + B_n x_n + C_n y_n \geq b_n\Big\}.\end{aligned} \tag{1}$$

Here, A_n, B_n, C_n, b_n denote appropriately defined data matrices and vectors. The sets X_n and Y_n comprise constraints only associated with x_n or y_n, *e.g.*, box constraints or non-negativity constraints. More precisely, we assume that both sets are intersections of polyhedral sets $\bar{X}_n, \bar{Y}_n$ and possible integrality constraints. In the following, we also refer to X_n as the *state space*.

An MS-MILP can then be expressed by its dynamic programming equations. For the root node, we obtain

$$v^* := \min_{x_r, y_r} \Big\{ f_r(x_r, y_r) + \sum_{m \in \mathcal{C}(r)} p_{rm} Q_m(x_r) \ : \ (x_r, y_r) \in \mathcal{F}_r(x_{a(r)}) \Big\} \tag{2}$$

with $x_{a(r)} = 0$, and v^* is the optimal value of the original problem. Let $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$. For all $n \in \overline{\mathcal{N}}$, the value function $Q_n : \mathbb{R}^{d_{a(n)}} \to \overline{\mathbb{R}}$ is defined by

$$Q_n(x_{a(n)}) := \min_{x_n, y_n} \Big\{ f_n(x_n, y_n) + \sum_{m \in \mathcal{C}(n)} p_{nm} Q_m(x_n) \ : \ (x_n, y_n) \in \mathcal{F}_n(x_{a(n)}) \Big\}.$$

For the leaf nodes $n \in \mathcal{N} \setminus \widetilde{\mathcal{N}}$, we set $\sum_{m \in \mathcal{C}(n)} p_{nm} Q_m(x_n) \equiv 0$. Moreover, we set $Q_n(x_{a(n)}) = +\infty$ if $\mathcal{F}_n(x_{a(n)}) = \emptyset$, and denote by $\mathrm{dom}(Q_n)$ the *effective domain* of $Q_n(\cdot)$.

Remark 2.1. *Note that regarding $Q_n(\cdot)$ as a function on $\mathbb{R}^{d_{a(n)}}$ is not standard in stochastic programming. Often it is (implicitly) assumed to be defined only on the domain $X_{a(n)}$. However, from our view, allowing $Q_n(\cdot)$ to be defined on $\mathbb{R}^{d_{a(n)}}$ with extended real values appears more suitable for the following steps.*

As this proves beneficial in the cut generation process, we introduce local variables z_n and accompany them with copy constraints $x_{a(n)} = z_n$ and constraints $z_n \in Z_{a(n)}$, with $Z_{a(n)} \supseteq X_{a(n)}$. The most natural choice is $Z_{a(n)} = X_{a(n)}$, but also other choices are possible, *e.g.*, $Z_{a(n)} = \mathrm{conv}(X_{a(n)})$. For more details we refer to [18]. This reformulation

yields the equivalent subproblems

$$Q_n(x_{a(n)}) = \min_{x_n, y_n, z_n} \Big\{ f_n(x_n, y_n) + \sum_{m \in \mathcal{C}(n)} p_{nm} Q_m(x_n) \ : (z_n, x_n, y_n) \in \mathcal{F}_n, \\ z_n = x_{a(n)},\ z_n \in Z_{a(n)} \Big\}, \tag{3}$$

where, $\mathcal{F}_n := \Big\{ (x_n, y_n, z_n) \in \mathbb{R}^{d_{a(n)}} \times \mathbb{R}^{d_n} \times \mathbb{R}^{\tilde{d}_n} \ : \ (x_n, y_n) \in \mathcal{F}_n(z_n) \Big\}$.

For the remainder of this article, we make some basic assumptions.

Assumption 1. *The following conditions are satisfied by* (1)-(3)*:*

(A1) For all $n \in \mathcal{N}$, the sets X_n and Y_n are compact.

(A2) For all $n \in \mathcal{N}$, all coefficients in A_n, B_n, C_n, b_n, f_n, $\bar{X}_n$ and $\bar{Y}_n$ are rational.

(A3) For all $n \in \overline{\mathcal{N}}$, $Z_{a(n)}$ is compact, rational MILP-representable and $Z_{a(n)} = \mathrm{dom}(Q_n)$.

Note that the boundedness in (A1) immediately implies that $F_n(x_{a(n)})$ is bounded for all $x_{a(n)} \in \mathbb{R}^{d_{a(n)}}$ and $n \in \mathcal{N}$. Property (A3) implies *relatively complete recourse*, which is a standard assumption in multistage stochastic programming. It allows us to focus on optimality cuts approximating $Q_n(\cdot)$ without requiring feasibility cuts that approximate $\mathrm{dom}(Q_n)$.

We obtain the following properties for the value functions.

Lemma 2.2. *Under Assumption 1, for all $n \in \overline{\mathcal{N}}$, the value functions $Q_n(\cdot)$ are proper, l.sc. (lower semicontinuous) and piecewise polyhedral with finitely many pieces.*

By applying the properness reasoning to the root node, we conclude that v^* is finite.

3 A New Cut Generation Framework

In this section, we present a novel framework for the generation of Lagrangian cuts in multistage stochastic mixed-integer programming, which serves as an alternative to the classical Lagrangian cut generation framework from SDDiP [31], which we state in Appendix A for comparison. The new framework is based on ideas that go back to [17], and in one specific form has been applied to two-stage stochastic programs in [11].

3.1 An Epigraph Perspective on Cut Generation

Recall the definition of the epigraph of the value functions $Q_n(\cdot), n \in \overline{\mathcal{N}}$:

$$\mathrm{epi}(Q_n) = \Big\{ (x_{a(n)}, \theta_n) \in \mathbb{R}^{d_{a(n)}} \times \mathbb{R} \ : \ \theta_n \geq Q_n(x_{a(n)}),\ x_{a(n)} \in \mathrm{dom}(Q_n) \Big\}. \tag{4}$$

We can use this definition to reformulate the subproblems (3) to

$$Q_n(x_{a(n)}) = \min_{x_n, y_n, z_n, (\theta_m)} \Big\{ f_n(x_n, y_n) + \sum_{m \in \mathcal{C}(n)} p_{nm} \theta_m \ : (z_n, x_n, y_n) \in \mathcal{F}_n, \\ z_n = x_{a(n)},\ z_n \in Z_{a(n)} \\ (x_n, \theta_m) \in \mathrm{epi}(Q_m),\ m \in \mathcal{C}(n) \Big\}. \tag{5}$$

Here, and in the following, we use (θ_m) as a shortened notation for $(\theta_m)_{m\in\mathcal{C}(n)}$.

Remark 3.1. *The condition $x_n \in \mathrm{dom}(Q_m)$ from the definition in (4) is always satisfied implicitly in problem (5), since $\mathrm{dom}(Q_m) = Z_n$ by Assumption 1 and $x_n \in X_n \subseteq Z_n$ by construction.*

In classical Benders-like decomposition methods, such as SDDiP, iteratively *optimality cuts* are constructed to approximate $Q_n(\cdot)$ and, if required, *feasibility cuts* are constructed to approximate $\mathrm{dom}(Q_n)$. This is done by solving distinct cut generation problems. However, from (4) it is evident that both types of cuts actually approximate $\mathrm{epi}(Q_n)$. Therefore, we may as well consider a unified cut generation problem to obtain polyhedral approximations Ψ_m of the sets $\mathrm{epi}(Q_m)$ [17]. Whereas we solely focus on optimality cuts in this paper, see Assumption 1, the resulting cut generation framework still proves itself valuable, as we shall see.

Remark 3.2. *In multistage stochastic programming the cuts are often aggregated to obtain cuts for the expected value functions $\sum_{m\in\mathcal{C}(n)} p_{nm}Q_m(x_n)$ (*single-cut *approach), as this reduces the total number of cuts in the subproblems. In this paper, instead, we consider a separate set of cuts,* i.e.*, separate approximations Ψ_m, for each $\mathrm{epi}(Q_m)$ (*multi-cut *approach). This approach is better suited to our cut generation framework.*

As the value functions $Q_n(\cdot)$ are not known explicitly within our decomposition method, we may not use subproblems (5) directly in the cut generation process. However, we can replace each occurrence of $\mathrm{epi}(Q_m)$ with its current approximation Ψ_m^{i+1} (with iteration index i). We refer to the associated value functions $\underline{Q}_m^{i+1}(\cdot)$ as *approximate value functions*. Note that this way, we actually generate cuts approximating $\mathrm{epi}(\underline{Q}_m^{i+1})$. However, by construction these cuts do also yield outer approximations of $\mathrm{epi}(\overline{Q}_m)$. To avoid unboundedness, each set is initialized with a valid outer approximation $\Psi_m^0, m \in \mathcal{C}(n)$.

For notational simplicity, for the remainder of this paper we define the set

$$\mathcal{W}_n^{i+1} := \Big\{ \big(x_n, y_n, z_n, (\theta_m)\big) \;:\; (x_n, y_n, z_n) \in \mathcal{F}_n,\ z_n \in Z_{a(n)},\ (x_n, \theta_m) \in \Psi_m^{i+1}, m \in \mathcal{C}(n) \Big\},$$

and further define $\lambda_n := \big(x_n, y_n, (\theta_m)\big)$ and $c_n^\top \lambda_n := f_n(x_n, y_n) + \sum_{m\in\mathcal{C}(n)} p_{nm}\theta_m$ (recall that $f(\cdot)$ is linear). Then, the approximate subproblems and approximate value functions associated with problem (5) can be compactly written as

$$\underline{Q}_n^{i+1}(x_{a(n)}^i) = \min_{\lambda_n, z_n} \left\{ c_n^\top \lambda_n \;:\; (\lambda_n, z_n) \in \mathcal{W}_n^{i+1},\ z_n = x_{a(n)}^i \right\}. \tag{6}$$

We make another assumption for the remainder of this paper.

Assumption 2. *For all $n \in \overline{\mathcal{N}}$ and all iterations i, all linear cuts defining the polyhedral set Ψ_m^{i+1} are defined by rational coefficients.*

Furthermore, in the next sections, we often require the convex hull $\mathrm{conv}(\mathcal{W}_n^{i+1})$ of a set $\mathcal{W}_n^{i+1}$. It has the following important property.

Remark 3.3. *Under Assumptions 1 and 2, the set $\mathrm{conv}(\mathcal{W}_n^{i+1})$ is a closed convex polyhedron. That means that there exist matrices $\tilde{A}_n, \tilde{B}_n$ and a vector $\tilde{d}_n$ such that*

$$\mathrm{conv}(\mathcal{W}_n^{i+1}) = \left\{ (\lambda_n, z_n) \;:\; \tilde{A}_n\lambda_n + \tilde{B}_n z_n \geq \tilde{d}_n \right\}.$$

3.2 A Feasibility Problem for the Epigraph

We can now start to address the actual cut generation process in the proposed unified framework. Given a point $(x^i_{a(n)}, \theta^i_n)$, we consider the feasibility problem

$$v^{f,i+1}_n(x^i_{a(n)}, \theta^i_n) := \min_{\lambda_n, z_n} \Big\{ 0 \ : \ (\lambda_n, z_n) \in \mathcal{W}^{i+1}_n, \ z_n = x^i_{a(n)}, \ \theta^i_n \geq c_n^\top \lambda_n \Big\}, \tag{7}$$

which can be shown to verify if $(x^i_{a(n)}, \theta^i_n) \in \operatorname{epi}(\underline{Q}^{i+1}_n)$.

Lemma 3.4. *Under Assumptions 1 and 2 and given a point $(x^i_{a(n)}, \theta^i_n)$, problem (7) is a feasibility problem for $\operatorname{epi}(\underline{Q}^{i+1}_n)$, that is,*

$$v^{f,i+1}_n(x^i_{a(n)}, \theta^i_n) = \begin{cases} 0, & \text{if } (x^i_{a(n)}, \theta^i_n) \in \operatorname{epi}(\underline{Q}^{i+1}_n) \\ +\infty, & \text{else.} \end{cases}$$

Proof. Let $(x^i_{a(n)}, \theta^i_n) \in \operatorname{epi}(\underline{Q}^{i+1}_n)$. Then according to (4) we have

$$\theta^i_n \geq \min_{\lambda_n, z_n} \Big\{ c_n^\top \lambda_n \ : \ (\lambda_n, z_n) \in \mathcal{W}^{i+1}_n, \ z_n = x^i_{a(n)} \Big\}. \tag{8}$$

This implies that there exists some (λ_n, z_n) such that for $(x^i_{a(n)}, \theta^i_n)$ all constraints of (7) are satisfied. Hence, $v^{f,i+1}_n(x^i_{a(n)}, \theta^i_n) = 0$.

Let $v^{f,i+1}_n(x^i_{a(n)}, \theta^i_n) = 0$. Then, there exist (λ_n, z_n) such that for $(x^i_{a(n)}, \theta^i_n)$ all constraints of (7) are satisfied. However, this implies (8), and thus $(x^i_{a(n)}, \theta^i_n) \in \operatorname{epi}(\underline{Q}^{i+1}_n)$. □

3.3 Lagrangian Cuts in the New Framework

To generate Lagrangian cuts, we apply a Lagrangian relaxation to problem (7). A key difference to the classical Lagrangian relaxation from SDDiP (see Appendix A) is that not only $x^i_{a(n)}$, but $(x^i_{a(n)}, \theta^i_n)$ is regarded as a fixed incumbent for the cut generation process. Therefore, we relax all constraints containing either $x^i_{a(n)}$ or θ^i_n. For given multipliers $(\pi_n, \pi_{n0}) \in \mathbb{R}^{d_{a(n)}} \times \mathbb{R}^+$ the dual function is then given by

$$\mathscr{L}^{i+1}_n(\pi_n, \pi_{n0}) := \min_{\lambda_n, z_n} \Big\{ \pi_n^\top z_n + \pi_{n0} c_n^\top \lambda_n \ : \ (\lambda_n, z_n) \in \mathcal{W}^{i+1}_n \Big\}. \tag{9}$$

The corresponding Lagrangian dual problem is

$$\max_{\pi_n, \pi_{n0}} \Big\{ \mathscr{L}^{i+1}_n(\pi_n, \pi_{n0}) - \pi_n^\top x^i_{a(n)} - \pi_{n0} \theta^i_n \ : \ \pi_{n0} \geq 0 \Big\}. \tag{10}$$

We state some important properties of this dual problem.

Theorem 3.5. *Under Assumptions 1 and 2, for the Lagrangian dual (10) it holds:*

(i) The dual function $\mathscr{L}^{i+1}_n(\cdot)$ is piecewise linear concave in (π_n, π_{n0}).

(ii) Its optimal value $\widehat{v}^{D,i+1}_n(x^i_{a(n)}, \theta^i_n)$ satisfies

$$\widehat{v}^{D,i+1}_n(x^i_{a(n)}, \theta^i_n) = \begin{cases} 0, & \text{if } (x^i_{a(n)}, \theta^i_n) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}^{i+1}_n)) \\ +\infty, & \text{else.} \end{cases}$$

Proof. (i) is a standard result on Lagrangian relaxation [see 20]. Another well-known property of the Lagrangian dual is that it is equivalent to a primal convexification of the original subproblem [19]. In our case, this convexification is given by

$$\min_{\lambda_n, z_n} \Big\{ 0 \ : \ (\lambda_n, z_n) \in \operatorname{conv}(\mathcal{W}_n^{i+1}),\ z_n = x^i_{a(n)},\ \theta^i_n \geq c_n^\top \lambda_n \Big\}. \tag{11}$$

The closed convex envelope $\overline{\operatorname{co}}(\underline{Q}_n^{i+1})(\cdot)$ can be expressed through the convex problem

$$\overline{\operatorname{co}}(\underline{Q}_n^{i+1})(x^i_{a(n)}) = \min_{\lambda_n, z_n} \Big\{ c_n^\top \lambda_n \ : \ (\lambda_n, z_n) \in \operatorname{conv}(\mathcal{W}_n^{i+1}),\ z_n = x^i_{a(n)} \Big\}, \tag{12}$$

see for instance [18, Theorems 3.8 and 3.9] for a formal proof. Therefore, by the same reasoning as in Lemma 3.4, problem (11) is a feasibility problem for $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$. □

Theorem 3.5 directly implies that the Lagrangian dual is unbounded whenever $(x^i_{a(n)}, \theta^i_n) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$. Therefore, there exists an unbounded ray (π^i_n, π^i_{n0}) such that

$$\mathscr{L}_n^{i+1}(\pi^i_n, \pi^i_{n0}) - (\pi^i_n)^\top x^i_{a(n)} - \pi^i_{n0}\theta^i_n > 0,$$

and $(x^i_{a(n)}, \theta^i_n)$ violates the following Lagrangian cut:

Definition 3.6. *For all $n \in \overline{\mathcal{N}}$ and some multipliers (π^i_n, π^i_{n0}), a Lagrangian cut is given by*

$$\pi^i_{n0}\theta_n + (\pi^i_n)^\top x_{a(n)} \geq \mathscr{L}_n^{i+1}(\pi^i_n, \pi^i_{n0}). \tag{13}$$

This type of cut is valid for any feasible (π^i_n, π^i_{n0}) in (10). We provide a proof in Appendix B.1.

Lemma 3.7. *Under Assumptions 1 and 2, for any $(\pi^i_n, \pi^i_{n0}) \in \mathbb{R}^{d_{a(n)}} \times \mathbb{R}^+$ the Lagrangian cut* (13) *is satisfied by all $(x_{a(n)}, \theta_n) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$, and thus by all $(x_{a(n)}, \theta_n) \in \operatorname{epi}(Q_n)$.*

We analyze the relation between the Lagrangian cuts (13) and the classical ones (22). We restrict to $\pi_{n0} > 0$ because $\pi_{n0} = 0$ leads to feasibility cuts that by assumption are not required in our case.

Remark 3.8. *Let $\pi^i_{n0} > 0$ in cut* (13)*. As shown in Proposition 1 in [11], with division by π^i_{n0}, it follows*

$$\begin{aligned}
\theta_n &\geq \frac{1}{\pi^i_{n0}} \mathscr{L}_n^{i+1}(\pi^i_n, \pi^i_{n0}) - \left(\frac{\pi^i_n}{\pi^i_{n0}}\right)^\top x_{a(n)} \\
&= \mathscr{L}_n^{i+1}(\widehat{\pi}_n, 1) - \widehat{\pi}_n^\top x_{a(n)} \\
&= \mathcal{L}_n^{i+1}(\widehat{\pi}_n) - \widehat{\pi}_n^\top x_{a(n)}
\end{aligned}$$

with $\widehat{\pi}_n := \frac{\pi^i_n}{\pi^i_{n0}}$. This is an equivalent representation of (13) *in the form of a classical Lagrangian optimality cut* (22)*, see Appendix A.*

We should emphasize that despite the scaling relation shown in Remark 3.8, the new cut generation framework may yield different cuts than the classical one because the choice of dual multipliers is based on a different dual problem.

3.4 Cut Selection Criteria

It is not immediately clear how to select cut coefficients (π_n^i, π_{n0}^i) from the unbounded Lagrangian dual (10) in the most reasonable way. On the one hand, computationally we aspire to determine coefficients (π_n^i, π_{n0}^i) by solving a bounded and feasible optimization problem instead of dealing with an unbounded one. On the other hand, we want to make sure that the obtained cuts are not only separating the incumbent $(x_{a(n)}^i, \theta_n^i)$ from $\text{epi}(Q_n)$, but also of good approximation quality. For instance, they should not be unnecessarily steep (see Sect. 1.1) and they should be supporting $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$.

The first aim can be achieved by bounding problem (10) artificially, *e.g.*, by introducing bounds on the dual multipliers. Another common approach is to fix its unbounded objective to 1. Combined with Remark 3.3 this allows to identify unbounded rays by analyzing a compact polyhedron called *alternative polyhedron* [17]. Finally, we may introduce a normalizing constraint to the Lagrangian dual (10).

With regard to the second aim, various quality criteria for cutting-planes have been put forward in the literature, see [14] for an overview. Many of these criteria have been applied in the context of BD or disjunctive programming before, as shown in Table 1, but our paper is the first one applying them to Lagrangian cuts, and incorporating most of them at once. We focus on three important criteria:

- *Facet-defining cuts.* These cuts reproduce facets of a convex polyhedral set, in our case $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$, and thus may be helpful in ensuring finite convergence.

- *Pareto-optimal cuts.* For $\pi_{n0} > 0$, Pareto-optimal cuts (13) are *non-dominated* in the sense that there exists no other cut

$$\tilde{\pi}_{n0}\theta_n + (\tilde{\pi}_n)^\top x_{a(n)} \geq \mathscr{L}_n^{i+1}(\tilde{\pi}_n, \tilde{\pi}_{n0})$$

  such that

$$\frac{\mathscr{L}_n^{i+1}(\tilde{\pi}_n, \tilde{\pi}_{n0}) - (\tilde{\pi}_n)^\top x_{a(n)}}{\tilde{\pi}_{n0}} \geq \frac{\mathscr{L}_n^{i+1}(\pi_n^i, \pi_{n0}^i) - (\pi_n^i)^\top x_{a(n)}}{\pi_{n0}^i}$$

  for all $(x_{a(n)}, \theta_n) \in \text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$. Note that Pareto-optimality with respect to $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ does not necessarily imply Pareto-optimality with respect to $\text{epi}(\underline{Q}_n^{i+1})$, but is easier to achieve. The idea of Pareto-optimal cuts was first put forward by [22].

- *Deep cuts.* The concept of deep cuts goes back to [8]. These cuts are *deep* in the sense that a maximum distance between the incumbent $(x_{a(n)}^i, \theta_n^i)$ and the separating hyperplane is realized, *i.e.*, they cut as deep as possible into the suboptimal region.

As shown in the literature, especially in [9], many of these criteria can be satisfied by optimizing over the so-called *reverse polar set* of $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ shifted by the incumbent $(x_{a(n)}^i, \theta_n^i)$. The reverse polar set is an important tool in the theory on cut generation, as it is directly linked to the support function of $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$, and thus provides a characterization of normal vectors of $(x_{a(n)}^i, \theta_n^i)$-separating hyperplanes [9, 13].

The reverse polar set was first introduced by [3] and can be defined as follows.

Table 1: Examination of cut quality criteria in the literature on BD and disjunctive programming.

Paper	MIS	Max. depth	Facet-def.	Pareto-opt.
Magnanti and Wong [22]				✓
Cornuéjols and Lemaréchal [13]		✓	✓	
Papadakos [23]				✓
Cadoux [10]		✓	✓	
Fischetti et al. [17]	✓			
Sherali and Lunday [27]				✓
Conforti and Wolsey [12]			✓	
Brandenberg and Stursberg [9]	✓		✓	✓
Hosseini and Turner [21]		✓		
Seo et al. [26]			✓	
This paper		✓	✓	✓

MIS: Cuts that correspond to minimal infeasible subsytems of the feasibility subproblem.

Definition 3.9. *The reverse polar set of a set $S \subset \mathbb{R}^n$ is defined as*

$$S^- := \Big\{ d \in \mathbb{R}^n \ : \ d^\top x \leq -1 \quad \forall x \in S \Big\}.$$

To simplify notation, we set $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n) := \big(\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1})) - (x^i_{a(n)}, \theta^i_n)\big)^-$ for the reverse polar set of $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ shifted by $(x^i_{a(n)}, \theta^i_n)$. Using Remark 3.3, it can be reformulated.

Lemma 3.10. *The reverse polar set $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$ can be expressed as*

$$\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n) = \left\{ (\gamma_n, \gamma_{n0}) \in \mathbb{R}^{d_{a(n)}} \times \mathbb{R} \ : \ \exists \mu_n \geq 0 \ : \ \begin{array}{l} \gamma_{n0} \leq 0 \\ -\tilde{A}_n^\top \mu_n - \gamma_{n0} c_n = 0 \\ -\tilde{B}_n^\top \mu_n - \gamma_n = 0 \\ \tilde{d}_n^\top \mu_n + \gamma_n^\top x^i_{a(n)} + \gamma_{n0} \theta^i_n \geq 1 \end{array} \right\}.$$

We provide a proof in Appendix B.2.

Remark 3.11. *Even with the above reformulation of $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$, an explicit formulation is usually not readily available due to the existence quantor and due to $\tilde{A}_n, \tilde{B}_n$ and $\tilde{d}_n$ not being known.*

Based on the existing work for BD, in the next sections, we present and investigate different strategies to generate Lagrangian cuts satisfying the above quality criteria. In the light of Remark 3.11, $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$ may not be used without further ado to generate such cuts computationally. Still, it proves useful in the derivation of Lagrangian cuts with favorable properties. In particular, as we shall see, optimizing over $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$ is closely linked to solving normalized Lagrangian dual problems. So in fact, our two perspectives to approach cut selection are intertwined and boil down to considering specific (bounded) normalizations of problem (10).

We first define the normalized Lagrangian dual in a general form and state some important properties. For simplicity, from now on, we assume that Assumptions 1 and 2

Table 2: Examination of normalized cut generation problems and different perspectives on it in the literature.

	Norm normalization			Linear normalization		
Paper	Prim	Proj	RP	Prim	Proj	RP
Cornuéjols and Lemaréchal [13]	✓	✓		✓	✓	✓
Cadoux [10]	✓	✓	✓			
Fischetti et al. [17]				(✓)		
Brandenberg and Stursberg [9]				✓		✓
Hosseini and Turner [21]	✓	✓		✓		
Chen and Luedtke [11]	(✓*)					
Seo et al. [26]					✓	
This paper	✓*	✓*	✓*	✓*	✓*	✓*

Prim: Primal perspective. Proj: Projection perspective. RP: Reverse polar perspective.
(✓): Perspective is applied, but not further explored. ✓*: Examination for Lagrangian cuts.

are satisfied, even if not explicitly stated.

Definition 3.12. *For some homogeneous normalization function $g_n : \mathbb{R}^{d_n} \times \mathbb{R}^+ \to \mathbb{R}$, the normalized Lagrangian dual is defined as*

$$\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n) := \max_{\pi_n, \pi_{n0}} \Big\{ \mathscr{L}_n^{i+1}(\pi_n, \pi_{n0}) - \pi_n^\top x^i_{a(n)} - \pi_{n0}\theta^i_n \ : \ g_n(\pi_n, \pi_{n0}) \le 1, \pi_{n0} \ge 0 \Big\}. \tag{14}$$

Remark 3.13. *As long as the normalization constraint $g_n(\pi_n, \pi_{n0}) \le 1$ is satisfied by some neighborhood N of the origin, we do not exclude any potential cuts due to the scaling property of π_{n0}, see Remark 3.8. In fact, Chen and Luedtke [11] prove that restricting the dual multipliers to $N \cap (\mathbb{R}^{d_{a(n)}} \times \mathbb{R}^+)$ yields a family of possible Lagrangian cuts that is satisfied by the same set of points $(x_{a(n)}, \theta_n)$ as the family of classical Lagrangian (optimality and feasibility) cuts from Appendix A.*

Lemma 3.14. *If $(x^i_{a(n)}, \theta^i_n) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$, then $\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n) = 0$, and vice versa.*

We provide a proof in Appendix B.3. Lemma 3.14 allows us to solely focus on the case where $(x^i_{a(n)}, \theta^i_n) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ for the remainder of this section.

In the next two subsections, we consider two different types of normalization: by norm constraints and by linear constraints. As we show, both types of normalization can be viewed from three different perspectives (a primal perspective, a projection perspective and a reverse polar perspective). These perspectives have been analyzed in the literature before, as shown in Table 2, but have not been linked all together in a generalized framework and have not been applied to Lagrangian cuts.

3.5 Normalization by Norm and Deep Cuts

We consider the normalized Lagrangian dual (14) and the associated Lagrangian cuts if some norm is used as the normalization function. We start by formally defining this type of cut.

Definition 3.15. *Let $\|\cdot\|$ be some arbitrary norm. The Lagrangian cut* (13) *defined by the solution (π_n, π_{n0}) to the normalized Lagrangian dual* (14) *with $g_n(\pi_n, \pi_{n0}) = \|\pi_n, \pi_{n0}\|$ is called* $\|\cdot\|$-deep Lagrangian cut*. For ℓ^p-norms we may also use the term* ℓ^p-deep Lagrangian cuts*.*

If appropriate norms are used, *e.g.*, the ℓ^1-norm or the ℓ^∞-norm, then the normalization can be expressed by linear constraints. However, due to the nonlinearity of $\mathscr{L}_n^{i+1}(\cdot)$, the cut generation subproblem is still not an LP, and therefore may be difficult to solve.

For the special choice of the ℓ^1-norm, this normalization is used by Chen and Luedtke [11] to generate Lagrangian cuts in two-stage stochastic programs, however without discussing the conceptual idea behind it in detail. Deep cuts allow for three theoretical and geometrical interpretations (cf. Table 2), which also explain why they are called *deep*. As the existing results from the literature can be applied to the multistage and Lagrangian setting in a straightforward way, we do not provide proofs here.

(1) **Maximizing cut depth.** Deep cuts maximize the distance between the incumbent $(x^i_{a(n)}, \theta^i_n)$ and the hyperplane associated with this cut in the dual norm $\|\cdot\|_*$ of $\|\cdot\|$, which means that they cut as deep as possible into the suboptimal region. Therefore, this distance can be interpreted as the *depth* or a scaled violation of this cut.

Lemma 3.16 (based on Hosseini and Turner [21])**.** *Let $\|\cdot\|$ be some norm and $\|\cdot\|_*$ its dual norm. Further, let $d_n\big((x^i_{a(n)}, \theta^i_n); (\pi_n, \pi_{n0})\big)$ denote the distance between the hyperplane defined by (π_n, π_{n0}) and the point $(x^i_{a(n)}, \theta^i_n) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ measured in $\|\cdot\|_*$. Then, the optimal value $\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n)$ of problem* (14) *with $g_n(\pi_n, \pi_{n0}) = \|\pi_n, \pi_{n0}\|$ equals*

$$\max_{\pi_n, \pi_{n0}} \Big\{ d_n\big((x^i_{a(n)}, \theta^i_n); (\pi_n, \pi_{n0})\big) \;:\; \pi_{n0} \geq 0 \Big\}.$$

(2) **Projection onto the epigraph.** From a primal perspective, generating $\|\cdot\|$-deep cuts is in some sense equivalent to minimizing the distance in $\|\cdot\|_*$ between the incumbent $(x^i_{a(n)}, \theta^i_n)$ and the epigraph $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$, *i.e.*, related to projecting $(x^i_{a(n)}, \theta^i_n)$ onto the epigraph.

Lemma 3.17 (based on Lemma 2.5 in [10])**.** *Let $\|\cdot\|$ be some norm and $\|\cdot\|_*$ its dual norm. Then, the optimal value $\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n)$ of problem* (14) *with $g_n(\pi_n, \pi_{n0}) = \|\pi_n, \pi_{n0}\|$ equals that of the projection problem*

$$\min_{x_{a(n)}, \theta_n} \Big\{ \|x_{a(n)} - x^i_{a(n)}, \theta_n - \theta^i_n\|_* \;:\; (x_{a(n)}, \theta_n) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1})) \Big\}. \tag{15}$$

Lemma 3.17 implies that $\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n) > 0$ if $(x^i_{a(n)}, \theta^i_n) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$, whereas $\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n) = 0$ if not, so it confirms Lemma 3.14. Therefore, as for the non-normalized case, we have a unique flag for cases where no separating cut has to be constructed. However, in contrast to the non-normalized case, the dual problem is bounded.

We can also conclude from Lemma 3.17 that a deep cut supports $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$.

Corollary 3.18 (based on Proposition 3 in Hosseini and Turner [21])**.** *Suppose $(x^i_{a(n)}, \theta^i_n) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ and let $(\widehat{x}_{a(n)}, \widehat{\theta}_{a(n)}) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ be a solution*

to (15). *Then, any* $\|\cdot\|$*-deep cut separating* $(x^i_{a(n)}, \theta^i_n)$ *from* $\mathrm{epi}(\overline{\mathrm{co}}(\underline{Q}^{i+1}_n))$ *supports* $\mathrm{epi}(\overline{\mathrm{co}}(\underline{Q}^{i+1}_n))$ *at* $(\widehat{x}_{a(n)}, \widehat{\theta}_{a(n)})$.

(3) **Minimizing a norm over the reverse polar set.** Interestingly, deep cuts allow for another geometric interpretation that is related to the reverse polar set $\mathcal{R}^{i+1}_n(x^i_{a(n)}, \theta^i_n)$. It is based on the observation that $\mathcal{R}^{i+1}_n(x^i_{a(n)}, \theta^i_n)$ is directly linked to the normals of separating hyperplanes.

Lemma 3.19 (Lemma 2.9 in [10])**.** *Let* (π^i_n, π^i_{n0}) *be the coefficients of a* $\|\cdot\|$*-deep cut constructed at* $(x^i_{a(n)}, \theta^i_n) \notin \mathrm{epi}(\overline{\mathrm{co}}(\underline{Q}^{i+1}_n))$. *Then there exists some* $\alpha > 0$ *such that* $-\alpha(\pi^i_n, \pi^i_{n0})$ *minimizes* $\|\cdot\|$ *over the reverse polar set* $\mathcal{R}^{i+1}_n(x^i_{a(n)}, \theta^i_n)$.

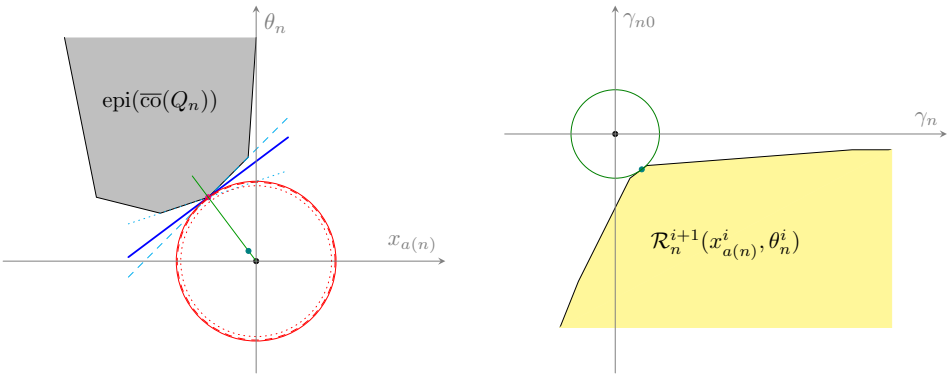
The main idea to prove this result is that due to positive homogeneity of norms, the norm $\|\cdot\|$ in the normalization constraint and the Lagrangian dual objective can be swapped.

To illustrate these three perspectives for different norms we provide an example, inspired by illustrations from the literature [9, 21].

Example 3.20. *We consider a given epigraph* $\mathrm{epi}(\overline{\mathrm{co}}(Q_n))$*, an incumbent* $(x^i_{a(n)}, \theta^i_n)$*, the associated reverse polar set* $\mathcal{R}^{i+1}_n(x^i_{a(n)}, \theta^i_n)$ *and the obtained deep Lagrangian cuts for different norms (*ℓ^2*,* ℓ^1*,* ℓ^∞ *and a weighted* ℓ^1*-norm). The sets and cuts are illustrated in Fig. 2-5 for the different norms. In each case, the illustration consists of two parts (a) and (b). In part (a), the incumbent (*black dot*) and the epigraph are depicted. Moreover, several norm balls are shown for the respective dual norms (*red lines*). We can see that the obtained deep Lagrangian cuts (*blue lines*) maximize the distance between the incumbent and the hyperplane in the dual norm. This is illustrated by depicting different valid cuts (*dashed/dotted cyan lines*) with smaller distances. On the other hand, it is also shown that the deep cuts minimize the distances between the incumbent and the epigraph in the dual norm and support the epigraphs at the corresponding projection to the epigraph (*violet line or point*). In part (b), the reverse polar set is depicted. Moreover, the optimal solutions (*teal line or point*) for optimizing the given norm (illustrated by a norm ball,* green line*) over the reverse polar set are highlighted. These solutions (apart from sign changes) characterize the normal vectors of the obtained cuts (see Lemma 3.19), as is additionally illustrated in part (a). Note that for none of the considered cases, the deep cuts are tight for* $\mathrm{epi}(\overline{\mathrm{co}}(Q_n))$ *at* $x^i_{a(n)}$*, contrary to classical Lagrangian cuts.*

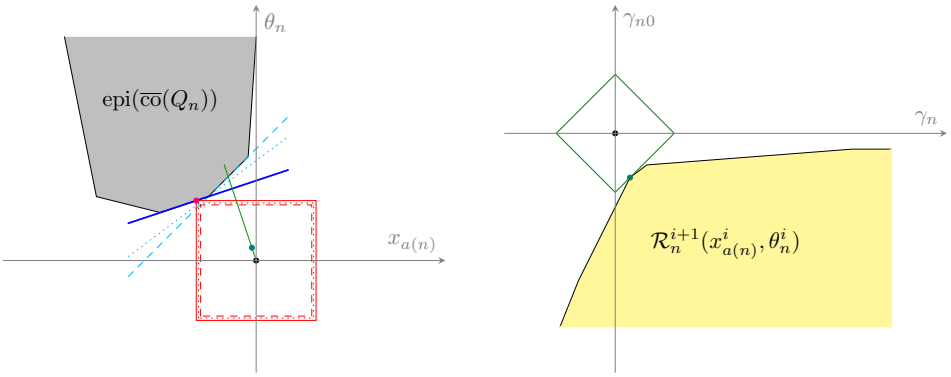
Example 3.20 also illustrates possible properties of deep cuts. Whereas deep cuts can be unique (see Fig. 2,3,4), also degenerate solutions with infinitely many different deep cuts are possible (see Fig. 5). This is the case if the optimization over $\mathcal{R}^{i+1}_n(x^i_{a(n)}, \theta^i_n)$ in Lemma 3.19 does not have a unique solution. Only for the ℓ^2-norm a unique solution is guaranteed. While non-unique solutions are not disadvantageous in general, degeneracy of the Lagrangian dual (14) may lead to selection of non-dominant cuts, compare Sect. 1.1.

Further, even if unique, deep cuts are not guaranteed to be facet-defining (see Fig. 2). In fact, while they cut deep into the suboptimal region, analyses in [10] (for the ℓ^2-norm) and in Hosseini and Turner [21] (for the ℓ^1-norm) show that they tend to be flat, at least in early stages of the algorithm when the optimality gap is still large. We can



(a) Epigraph and deep cut. (b) Reverse polar set.

Figure 2: Illustration of deep cuts for the ℓ^2-norm.



(a) Epigraph and deep cut. (b) Reverse polar set.

Figure 3: Illustration of deep cuts for the ℓ^1-norm.

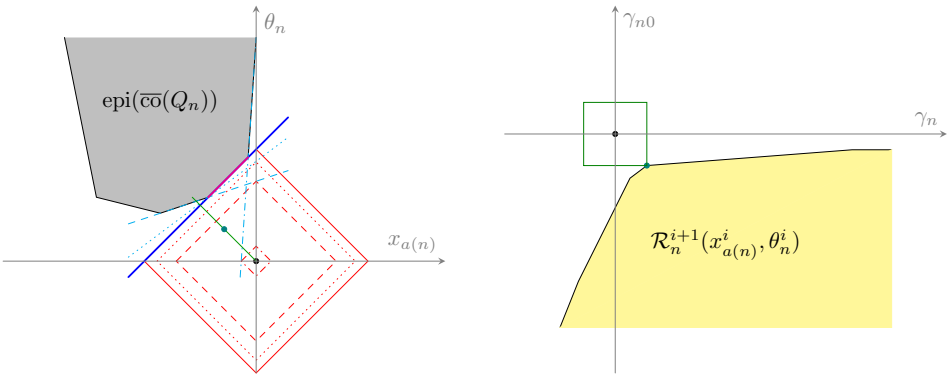
anticipate from Fig. 3 and Lemma 3.19, though, that deep cuts are facet-defining if $\|\cdot\|_*$ takes its minimum over $\mathcal{R}_n(x_{a(n)}^i, \theta_n^i)$ in a vertex.

Finally, also the projection problem (15) is not guaranteed to have a unique solution for all but the ℓ^2-norm (see Fig. 4). If the solution is non-unique, then the associated deep cut is unique and facet-defining.

Remark 3.21. *Another intuitive way to bound the Lagrangian dual* (10) *and to select cut coefficients* (π_n, π_{n0}) *is to introduce simple box constraints for the multipliers. For symmetric boxes around the origin, which can be modeled by a weighted* ℓ^∞*-norm, we may interpret the obtained cuts as deep cuts. However, which specific cuts are selected highly depends on the chosen multiplier bounds. Too large bounds may favor degeneracy and very steep cuts, for too small bounds, only almost horizontal cuts can be selected.*

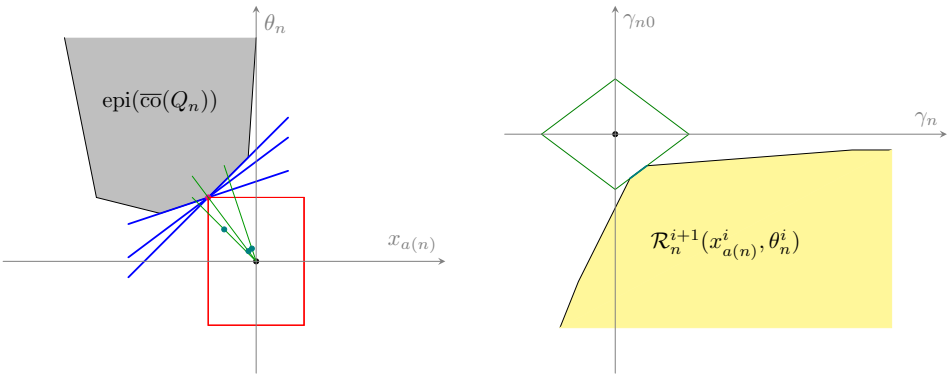
3.6 Linear Normalization

We consider the normalized Lagrangian dual (14) with a linear normalization function $g_n(\pi_n, \pi_{n0}) = u_n^\top \pi_n + u_{n0}\pi_{n0}$ defined by some coefficients $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \in \mathbb{R}$. Recall that the initial Lagrangian dual problem (10) is unbounded and that we introduce the normalization constraint in (14) in order to transform the problem to a bounded one to identify unbounded rays. In contrast to the norm-based normalization from Sect. 3.5,



(a) Epigraph and deep cut. (b) Reverse polar set.

Figure 4: Illustration of deep cuts for the ℓ^∞-norm.



(a) Epigraph and deep cut. (b) Reverse polar set.

Figure 5: Illustration of deep cuts for a weighted ℓ^1-norm.

a linear normalization does not guarantee boundedness, though. Hence, the choice of (u_n, u_{n0}) is crucial to ensure that an optimal solution exists. We further analyze this later in this section, but for now take the following assumption.

Assumption 3. *Given some $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$, the normalized Lagrangian dual* (14) *with $g_n(\pi_n, \pi_{n0}) = u_n^\top \pi_n + u_{n0}\pi_{n0}$ has a finite optimal value $\widehat{v}_n^{ND,i+1}(x^i_{a(n)}, \theta^i_n) > 0$.*

3.6.1 Linear Normalization Cuts

We can then define the associated type of Lagrangian cuts.

Definition 3.22. *Let $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$ and let the normalized Lagrangian dual* (14) *satisfy Assumption 3. Then, we refer to the Lagrangian cut* (13) *defined by its solution (π_n, π_{n0}) as a* linear normalization (LN) Lagrangian cut.

Again, we can take three different perspectives on LN cuts.

(1) **Pseudonorm perspective.** Hosseini and Turner [21] restrict to choices of (u_n, u_{n0}) such that $g_n(\pi_n, \pi_{n0}) = u_n^\top \pi_n + u_{n0}\pi_{n0} \geq 0$ for all $(\pi_n, \pi_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}_+$. In such a case, $g_n(\cdot)$ is a *linear pseudonorm* (in contrast to norms, it is not positive definite). This means that LN cuts can be interpreted as maximizing the distance between the associated hyperplane and $(\widehat{x}^i_{a(n)}, \widehat{\theta}^i_n)$ in a linear pseudonorm.

(2) **Projection on a line segment.** This perspective has been brought up several times in the literature in different variants, most recently by Seo et al. [26]. Even though the geometric idea is the same in the Lagrangian context, the formal description changes a bit.

First, we exploit that for linear $g_n(\cdot)$, the normalized Lagrangian dual (14) can be reformulated as an LP and then be dualized with strong duality.

Lemma 3.23. *Let $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$. The normalized Lagrangian dual (14) with $g_n(\pi_n, \pi_{n0}) = u_n^\top \pi_n + u_{n0}\pi_{n0}$ can be formulated as an LP, and its dual is*

$$\min_{\lambda_n, z_n, \eta_n} \Big\{ \eta_n \ : \ (\lambda_n, z_n) \in \operatorname{conv}(\mathcal{W}_n^{i+1}), \quad \eta_n \geq 0, \ u_{n0}\eta_n \geq c_n^\top \lambda_n - \theta_n^i, \\ u_n \eta_n = z_n - x_{a(n)}^i \Big\}. \tag{16}$$

We provide a proof in Appendix B.4. Problem (16) can be interpreted as finding the smallest scaling factor $\eta_n \geq 0$ such that starting from $(x_{a(n)}^i, \theta_n^i)$ along direction (u_n, u_{n0}) a point in $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ is reached. Again, for the optimal value it follows $\widehat{v}_n^{ND,i+1}(x_{a(n)}^i, \theta_n^i) = \eta_n^* > 0$ if and only if $(x_{a(n)}^i, \theta_n^i) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$. Given the optimal value, the projection of $(x_{a(n)}^i, \theta_n^i)$ onto $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ along direction (u_n, u_{n0}) can be determined as

$$(\widehat{x}_{a(n)}, \widehat{\theta}_{a(n)}) = (x_{a(n)}^i, \theta_n^i) + \eta_n^*(u_n, u_{n0}). \tag{17}$$

We then obtain the following result:

Corollary 3.24. *Let $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$ and let the normalized Lagrangian dual (14) satisfy Assumption 3. Furthermore, let $(\widehat{x}_{a(n)}, \widehat{\theta}_{a(n)}) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ satisfy (17). Then, the associated LN cut supports $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ at $(\widehat{x}_{a(n)}, \widehat{\theta}_{a(n)})$.*

(3) **Maximizing a linear function over the reverse polar set.** Again, an interpretation with respect to $\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i)$ is possible based on its characterization of normal vectors of separating hyperplanes. More precisely, LN cuts can be obtained by maximizing the linear objective function $u_n^\top \gamma_n + u_{n0}\gamma_{n0}$ over $\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i)$:

$$\max_{\gamma_n, \gamma_{n0}} \Big\{ u_n^\top \gamma_n + u_{n0}\gamma_{n0} \ : \ (\gamma_n, \gamma_{n0}) \in \mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i) \Big\}. \tag{18}$$

This perspective is discussed in detail in Brandenberg and Stursberg [9] for BD and in [13] for general convex sets (where solving problem (18) is shown to correspond to evaluating a function called *reverse gauge*). For the relation to the normalized Lagrangian dual (14), the following result holds. For a sketch of the proof, see Appendix B.5.

Theorem 3.25 (based on [9])**.** *Let $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$ and $(x_{a(n)}^i, \theta_n^i) \notin \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$.*

(i) If the normalized Lagrangian dual (14) satisfies Assumption 3, then problem (18) has a finite optimal value, and vice versa. The optimal points are the same up to scaling with some negative scalar and the optimal values multiply to -1.

(ii) The induced cuts for $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ are equivalent for both problems.

Geometrically, a favorable property of generating cuts based on $\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i)$ is that supporting cuts for $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ can be obtained in a straightforward way. Even more, if $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ is a full-dimensional polyhedron, then each vertex (γ_n, γ_{n0}) of $\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i)$ corresponds to the normal vector of a facet of $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$, and vice versa [9]. In order to identify such points, we can use the LP (18), given a choice of $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$ such that a finite optimum is attained.

In fact, perspectives (2) and (3) make a sufficient condition for such a finite optimum readily available, and by that also allow us to conclude when Assumption 3 is satisfied. This result is already proven in [13] and [9] using perspective (3), and then follows for the normalized Lagrangian dual (14) with Theorem 3.25. In Appendix B.6, we provide an alternative proof based on perspective (2).

Lemma 3.26. *Assumption 3 is satisfied if*

$$(u_n, u_{n0}) \in \text{cone}\left(\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1})) - (x_{a(n)}^i, \theta_n^i)\right) \setminus \{0\}, \tag{19}$$

where $\text{cone}(S)$ *denotes the conical hull of a set* S*.*

Lemma 3.26 implies that also choosing (u_n, u_{n0}) from $\text{epi}(\underline{Q}_n^{i+1}) - (x_{a(n)}^i, \theta_n^i)$ or even from $\text{epi}(Q_n) - (x_{a(n)}^i, \theta_n^i)$ is sufficient. In other words, choosing reasonable coefficients $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$ boils down to finding a *core point* within one of these epigraphs. We discuss this in more detail in Sect. 3.6.2.

We can now address some beneficial properties of LN cuts with respect to the aforementioned cut quality criteria. According to Theorem 3.25 there exists a one-to-one relation between optimal solutions of the normalized Lagrangian dual (14) and problem (18). However, this does not extend to optimal vertices, leading to a slightly less strong result for facet-defining cuts with respect to problem (14) [9]. We obtain the following properties:

Theorem 3.27. *Let* $(u_n, u_{n0}) \in \mathbb{R}^{d_n} \times \mathbb{R}$*. Consider the normalized Lagrangian dual problem* (14) *with* $g_n(\pi_n, \pi_{n0}) = u_n^\top \pi_n + u_{n0}\pi_{n0}$*. Then,*

(i) for all $(u_n, u_{n0}) \in \text{cone}\left(\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1})) - (x_{a(n)}^i, \theta_n^i)\right) \setminus \{0\}$*, the optimal point* (π_n^*, π_{n0}^*) *defines a supporting cut for* $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$*,*

(ii) for all $(u_n, u_{n0}) \in \text{cone}\left(\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1})) - (x_{a(n)}^i, \theta_n^i)\right) \setminus \{0\}$*, there exists an optimal extreme point* (π_n^*, π_{n0}^*)*, such that the obtained cut is facet-defining for* $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$*,*

(iii) for all $(u_n, u_{n0}) \in \text{relint}\left(\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1})) - (x_{a(n)}^i, \theta_n^i)\right)$*, any optimal point* (π_n^*, π_{n0}^*) *with* $\pi_{n0}^* > 0$ *defines a Pareto-optimal cut for* $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ *on* $\text{conv}(Z_{a(n)})$*.*

Part (i) directly follows from Lemma 3.26 and Corollary 3.24. Part (ii) follows from Theorem 3.3 in [9], and part (iii) follows from Theorem 3.43 in [28]. Note that the results from the literature require to choose (u_n, u_{n0}) from the relative interior of the epigraph restricted to $\text{conv}(X_{a(n)})$. However, under Assumption 1, if we choose $Z_n = X_{a(n)}$ or $Z_n = \text{conv}(X_{a(n)})$, in our case the considered epigraphs are always restricted to this set.

Considering part (ii), the only case in which no facet-defining cut is obtained occurs if the optimal solution to the normalized Lagrangian dual (14) is not unique. However, this is only the case for a small subset of choices (u_n, u_{n0}). Especially if the choice is adapted in each iteration, the occurrences of such cases should be negligible [9]. With

respect to (iii), we should emphasize again that Pareto-optimality for $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ does not necessarily imply Pareto-optimality for $\text{epi}(\underline{Q}_n^{i+1})$.

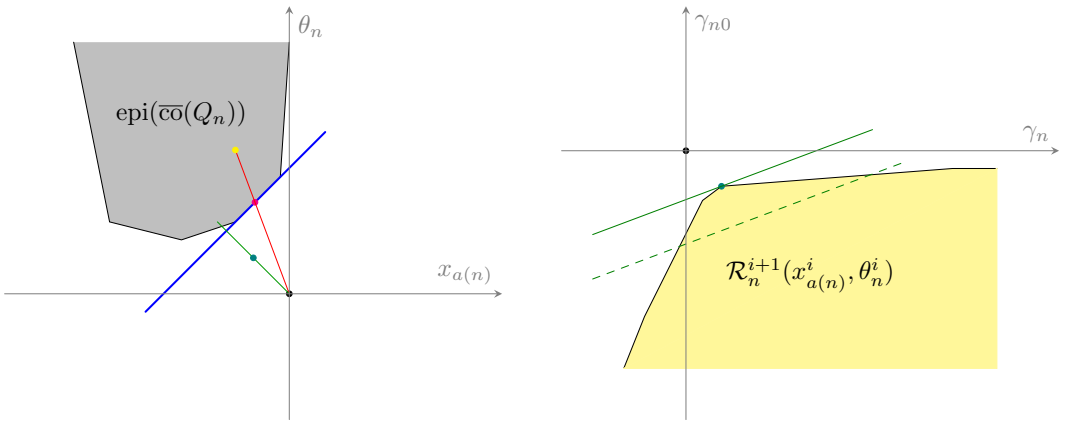
We finish our theoretical results in this subsection with two remarks.

Remark 3.28. *Based on the perspective taken, LN cuts are also called* pseudo-deep cuts *(perspective (1), Hosseini and Turner [21]) or* closest cuts *(perspective (2), Seo et al. [26]) in the literature. We could also refer to them as* core point cuts*, or based on perspective (3) as* reverse gauge cuts*.*

Remark 3.29. *Choosing $(u_n, u_{n0}) = (0, 1)$ in Definition 3.22 yields the classical Lagrangian cuts presented in Appendix A.*

We illustrate the different perspectives on LN cuts again using an example.

Example 3.30. *Consider epigraph $\text{epi}(\overline{\text{co}}(Q_n))$, incumbent $(x^i_{a(n)}, \theta^i_n)$ and reverse polar set $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$ from Example 3.30. These objects and an exemplary LN cut are illustrated in Fig. 6. In part (a) the geometric idea of projection along a line segment is highlighted. The direction of this line segment is (u_n, u_{n0}) and obtained as the difference between a known core point (*yellow dot*) in $\text{epi}(\overline{\text{co}}(Q_n))$ and $(x^i_{a(n)}, \theta^i_n)$. In part (b) we can see that (apart from sign changes) the cut normal to the LN cut can be determined by maximizing the linear function $u_n^\top \gamma_n + u_{n0}\gamma_{n0}$ over $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$. As the solution is an extreme point of $\mathcal{R}_n^{i+1}(x^i_{a(n)}, \theta^i_n)$ (*green dot*), the corresponding LN cut is facet-defining.*



(a) Epigraph and deep cut.

(b) Reverse polar set.

Figure 6: Illustration of LN cuts given some known core point.

3.6.2 Identifying Core Points

As described before, a key challenge of generating LN cuts with favorable properties is to choose (u_n, u_{n0}) appropriately, *i.e.*, according to Lemma 3.26 or Theorem 3.27. In the literature on decomposition methods, it is often proposed to evaluate feasible points in the objective function to obtain core points. Whereas this approach is straightforward for BD or the two-stage stochastic case [9, 28], in the multistage case, evaluating $Q_n(\cdot)$ exactly is computationally prohibitive in general. Furthermore, note that we also cannot evaluate $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$ efficiently, as this function is not known explicitly and evaluating it requires to solve a Lagrangian dual problem. Therefore, we propose heuristics to

identify potential core points $(\widehat{x}^i_{a(n)}, \widehat{\theta}^i_n)$, and by that reasonable coefficients (u_n, u_{n0}), based on function $\underline{Q}^{i+1}_n(\cdot)$. This comes with some considerable challenges, as we shall see. Some of these heuristics are particularly suited to SDDiP where $X_{a(n)} = \{0,1\}^{d_{a(n)}}$.

We consider the following approaches:

- `Mid`. We set $\widehat{x}^i_{a(n)} = \text{mid}(\text{conv}(X_{a(n)}))$ where mid denotes the midpoint (we assume box constraints for $x_{a(n)}$). The idea is that incentivizing the LN cuts to support $\text{epi}(\overline{\text{co}}(\underline{Q}^{i+1}_n)$ in the interior of the state space may be useful to avoid the degeneracy issues for SDDiP discussed in Sect. 1.1. We then evaluate the approximate value function to obtain $\widehat{\theta}^i_n = \underline{Q}^{i+1}_n(\widehat{x}^i_{a(n)})$, even if this does not satisfy the relative interior requirement in Theorem 3.27.

- `In-Out`. We use the dynamic approach to compute core points proposed by Papadakos [23] heuristically, that is, we set $\widehat{x}^i_{a(n)} = \frac{1}{2}\widehat{x}^{i-1}_{a(n)} + \frac{1}{2}x^i_{a(n)}$ and $\widehat{\theta}^i_n = \underline{Q}^{i+1}_n(\widehat{x}^i_{a(n)})$ to obtain a candidate core point.

- `Eps`. For some $\varepsilon > 0$, we use an ε-perturbation of $x^i_{a(n)}$ into the interior of $\text{conv}(X_{a(n)})$, and set $\widehat{\theta}^i_n = \underline{Q}^{i+1}_n(\widehat{x}^i_{a(n)})$. This idea is inspired by the perturbation strategy described in Sherali and Lunday [27]. A similar approach is also used in Seo et al. [26]. It is particularly suited to SDDiP, as it avoids the possible degeneracy issues related to generating cuts at extreme points of the state space, see Sect. 1.1, and thus may contribute to identifying facet-defining cuts. For sufficiently small ε, the LN cut should still be supporting $\text{epi}(\overline{\text{co}}(\underline{Q}^{i+1}_n))$ at $(x^i_{a(n)}, \overline{\text{co}}(\underline{Q}^{i+1}_n)(x^i_{a(n)}))$.

- `Relint`. We solve an auxiliary feasibility problem with slack variables to find a potential core point in relint $\big(\text{epi}(\underline{Q}^{i+1}_n)\big)$. This feasibility problem is defined in a similar way to the one described in Sect. 5.1 of Conforti and Wolsey [12].

These heuristics are illustrated in Fig. 7 for a simple example, where all of them are sufficient to get reasonable directions (u_n, u_{n0}), but lead to very different cuts being selected.

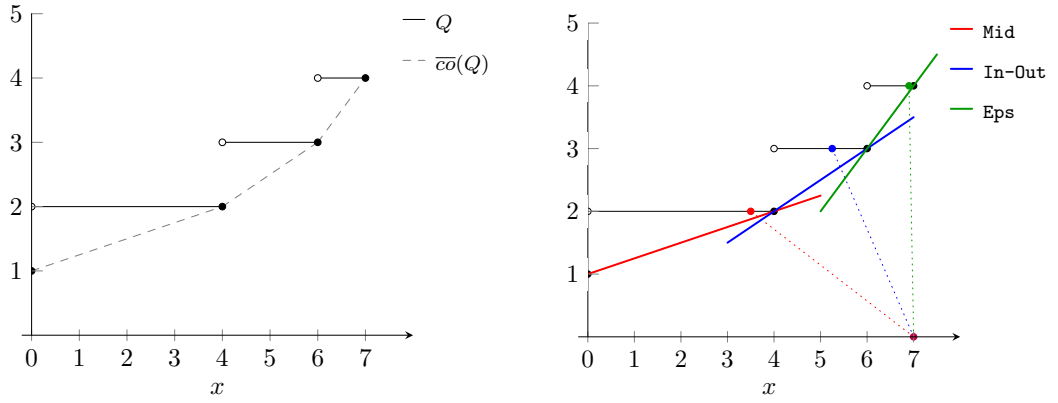


Figure 7: Illustration of core point identification heuristics.

Note. LEFT: Value function $Q(\cdot)$ and convex envelope $\overline{\text{co}}(Q)(\cdot)$ for an example with continuous state space $X = [0, 7]$. RIGHT: Identification of different core points and computation of the associated LN cuts for incumbent $(\bar{x}, \bar{\theta}) = (7, 0)$. For `In-Out` we assume $\widehat{x}^{i-1} = \frac{7}{2}$.

Despite their straightforwardness, the aforementioned approaches come with some notable challenges. Crucially, the first three approaches yield candidates satisfying $\widehat{x}^i_{a(n)} \in \operatorname{conv}(X_{a(n)})$. However, this choice is not necessarily feasible if integrality constraints are present, thus leading to $\underline{Q}_n^{i+1}(\widehat{x}^i_{a(n)}) = +\infty$ (especially if $Z_{a(n)} = X_{a(n)}$). We thus do not obtain a core point or a reasonable choice of (u_n, u_{n0}). This is illustrated in Fig. 8. Instead of evaluating $\underline{Q}_n^{i+1}(\cdot)$, in such a case we may revert to the value function $\underline{Q}_n^{LP,i+1}(\cdot)$ of the associated LP relaxation (`LP-value`). While this may yield a sufficient direction (u_n, u_{n0}), it may also yield one that does not point into $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$, see Fig. 8. To mitigate this risk, we may alternatively utilize the current state θ_n^i (`epi-state`) or the primal objective value $\underline{Q}_n^{i+1}(x^i_{a(n)})$ at the incumbent (`primal-obj`) if they exceed $\underline{Q}_n^{LP,i+1}(\widehat{x}^i_{a(n)})$, see again Fig. 8. All these approaches have no guarantees to yield true core points, though.

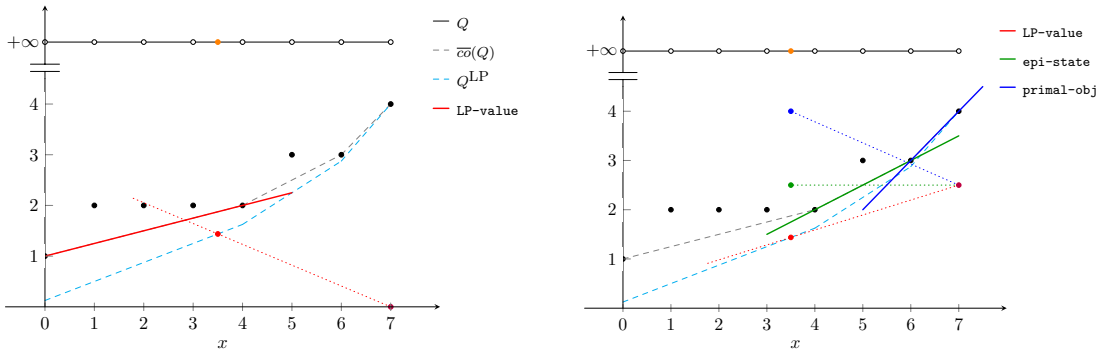


Figure 8: Core point identification challenges for integer state space.

Note. LEFT: Value function $Q(\cdot)$ and convex envelope $\overline{\text{co}}(Q)(\cdot)$ for an example with integer state space $X = \{0, 1, 2, \ldots, 7\}$. Given incumbent $(\bar{x}, \bar{\theta}) = (7, 0)$, using the LP relaxation value leads to a sufficient core point candidate and generation of an LN cut. RIGHT: The approach `LP-value` is not sufficient given incumbent $(\bar{x}, \bar{\theta}) = (7, \frac{5}{2})$. A sufficient core point candidate for generation of an LN cut is obtained using `epi-state` or `primal-obj`, though.

Even with the proposed heuristics we may end up with a normalized Lagrangian dual problem (14) that is unbounded. This is unintended, because the decomposition method may terminate with an error. Therefore, we should check for a potential unboundedness in practice, and if it is detected, take some special counter-measures. For instance, we could try another heuristic, generate a different type of cut instead of an LN cut, *e.g.*, a strengthened Benders cuts, or artificially bound the normalized Lagrangian dual problem (14).

Problem (16) provides a natural way to check for unboundedness, or the validity of direction (u_n, u_{n0}), respectively. However, it cannot be solved immediately, as we do not know $\operatorname{conv}(\mathcal{W}_n^{i+1})$ explicitly. We may instead solve the approximation

$$\min_{\lambda_n, z_n, \eta_n} \left\{ \eta_n \;:\; (\lambda_n, z_n) \in \mathcal{W}_n^{i+1},\ \eta_n \geq 0,\ u_{n0}\eta_n \geq c_n^\top \lambda_n - \theta_n^i,\ u_n \eta_n = z_n - x^i_{a(n)} \right\} \tag{20}$$

using the known set $\mathcal{W}_n^{i+1}$ instead of $\operatorname{conv}(\mathcal{W}_n^{i+1})$. If the normalized Lagrangian dual (14) is unbounded, then this problem is infeasible. Therefore, we can use infeasibility of (20) as a proxy for possible unboundedness. Unfortunately, in the presence of integrality constraints, infeasibility of (20) may occur very often, even given an appropriate direc-

tion (u_n, u_{n0}), thus leading to taking the previously discussed counter-measures more often than required. On the other hand, using the LP relaxation of (20) is not sufficient to rule out all cases of unboundedness. For an effective practical implementation of LN cuts this is a significant challenge.

Finally, let us present an alternative, fifth approach to come up with core points.

- `Conv.` We note that any convex combination of two (or more) feasible points $\left(x_1, \underline{Q}_n^{i+1}(x_1)\right)$ and $\left(x_2, \underline{Q}_n^{i+1}(x_2)\right)$ is always contained in $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$. One such point is readily available with $\left(x^i_{a(n)}, \underline{Q}_n^{i+1}(x^i_{a(n)})\right)$. For the case $X_{a(n)} = \{0,1\}^{d_{a(n)}}$, an intuitive strategy to obtain a second one is to consider the diagonal counterpart of $x^i_{a(n)}$ (swapping 0 and 1 for all components) and its function value for $\underline{Q}_n^{i+1}(\cdot)$. If this counterpart is feasible, we obtain a whole family of core points. Setting the convex combination parameter appropriately, we may even obtain a core point with first component $\text{mid}(\text{conv}(X_{a(n)}))$, but without having to evaluate the approximate value function in a non-integer state. This approach is highlighted in Fig. 9.

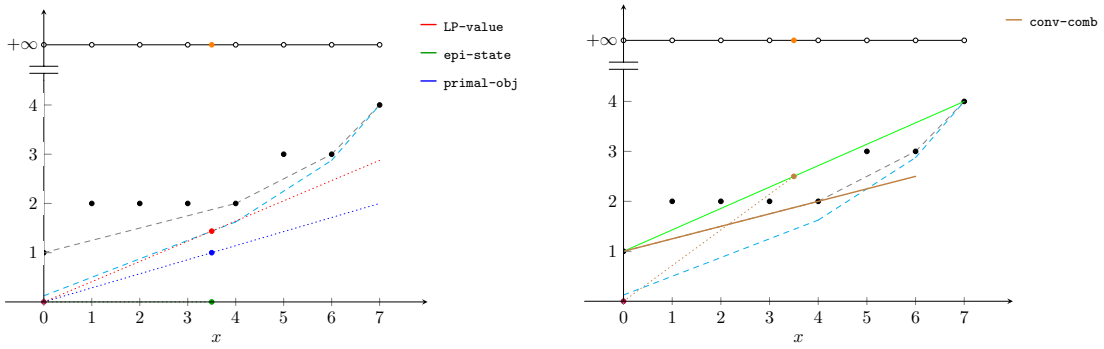


Figure 9: Illustration of core point identification heuristics.

Note. LEFT: Neither `LP-Value`, `epi-state` nor `primal-obj` are sufficient to identify an appropriate projection direction given incumbent $(\bar{x}, \bar{\theta}) = (0, 0)$. RIGHT: Choosing a core point as a convex combination of $(0, 1)$ and $(7, 4)$ (light green) with parameter $\frac{1}{2}$ and the resulting LN cut.

4 Convergence of NBD and SDDiP

For the case of binary state variables, *i.e.*, $x_n \in \{0,1\}^{d_n}$ for all $n \in \mathcal{N}$, classical Lagrangian cuts are sufficient to ensure (almost sure) finite convergence of decomposition methods such as NBD or SDDiP because these cuts are valid, tight and finite (see Appendix A). In this section, we address the convergence properties of these methods if the proposed new cut generation framework is incorporated. We restrict our discussion to NBD, that is exploring the whole scenario tree in each iteration. The considered NBD algorithm is displayed in Appendix C.

For SDDiP, only a sample of scenarios is considered in each iteration, so we have to replace line 6 of the NBD algorithm in Appendix C with a sampling step, but stagewise independence is assumed. Given the convergence of NBD, the almost sure finite convergence of SDDiP, follows with the same arguments as in the original SDDiP article [31, Theorem 2]. Moreover, instead of presenting a complete, self-contained convergence proof, we focus on the decisive properties of the Lagrangian cuts obtained in

our new cut generation framework. The rest of the convergence proof remains unchanged compared to [31].

The validity of the new Lagrangian cuts is already proven in Lemma 3.7. Therefore, some results with respect to finiteness and tightness remain to be proven. An important property in that regard is the polyhedrality of the closed convex envelopes $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$.

Lemma 4.1. *For all $n \in \overline{\mathcal{N}}$ and all iterations i, $\overline{\text{co}}(Q_n)(\cdot)$ and $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$ are piecewise linear convex functions.*

Lemma 4.1 is proven in Appendix B.7. It implies that $\text{epi}(\overline{\text{co}}(Q_n))$ and $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ are polyhedra with finitely many facets. Additionally, we take a technical assumption:

Assumption 4. *In any node $n \in \overline{\mathcal{N}}$ and any iteration i of NBD, given the same trial point $(x_{a(n)}^i, \theta_n^i)$ and the same approximations Ψ_m^{i+1} for all $m \in \mathcal{C}(n)$, solving the normalized Lagrangian dual problem* (14) *yields the same cut.*

This assumption is required to avoid that infinitely many different cuts are generated given the same state and the same approximate value function. It should be satisfied by most deterministic MILP solvers.

4.1 Results for LN Cuts

For NBD using LN cuts, we can exploit that the cuts are guaranteed to be (almost always) facet-defining. If we restrict to facet-defining cuts, we obtain the following convergence result.

Theorem 4.2. *Let $X_n = \{0,1\}^{d_n}$ for all $n \in \overline{\mathcal{N}}$. Assume that for any $n \in \overline{\mathcal{N}}$ and any iteration i the normalized Lagrangian dual problem* (14) *and the generated cuts satisfy Theorem 3.27 (ii). Then, after a finite number of iterations i, for all $n \in \overline{\mathcal{N}}$ the approximations Ψ_n^{i+1} are exact for $\text{epi}(Q_n(\cdot))$ at all $x_n^i \in X_{a(n)}$ computed in the forward pass.*

We provide a proof in Appendix B.8. Note that the proof still works if the number of steps between the generation of facet-defining cuts is always finite.

4.2 Results for Deep Cuts

For deep Lagrangian cuts proving convergence is a bit more tedious, since the facet property is not as straightforward to assure. The main idea to prove finite convergence thus is the following: Even if deep cuts are not guaranteed to be facet-defining or tight at $x_{a(n)}^i$ in general, such tight cuts will eventually be generated after finitely many steps. This is illustrated in Fig. 10.

As a first step, we introduce an auxiliary result where we consider a sequence of trial points with fixed first component $\bar{x}_{a(n)}$. Here, we do not exploit the binary nature of the state variables yet, but only the polyhedrality of $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$. For notational convenience, we set $\bar{\theta}_n := \overline{\text{co}}(\underline{Q}_n^{i+1})(\bar{x}_{a(n)})$ for the remainder of this section. The proof is shown in Appendix B.9.

Lemma 4.3. *For any $n \in \overline{\mathcal{N}}$, consider a subsequence of trial points $(\bar{x}_{a(n)}, \theta_n^{i_\nu})_{\nu \in \mathbb{N}}$ with fixed first component $\bar{x}_{a(n)} \in X_{a(n)}$ for all $\nu \in \mathbb{N}$. Then, the point $(\bar{x}_{a(n)}, \bar{\theta}_n)$ and any sequence $(\widehat{x}_{a(n)}^{i_\nu}, \widehat{\theta}_n^{i_\nu})_{\nu \in \mathbb{N}}$ of solutions of the projection problems* (15) *satisfy $\lim_{\nu \to \infty}(\widehat{x}_{a(n)}^{i_\nu}, \widehat{\theta}_n^{i_\nu}) = (\bar{x}_{a(n)}, \bar{\theta}_n)$.*

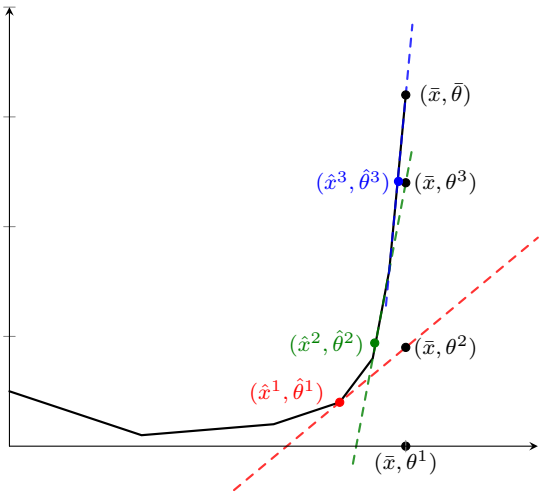


Figure 10: Illustration of deep cuts leading to tightness after finitely many steps for fixed $\bar{x}$. The blue cut is tight at $(\bar{x}, \bar{\theta})$, thus equal to a classical Lagrangian cut from Appendix A.

We now consider subproblems with fixed first component $\bar{x}_{a(n)}$ and with fixed approximations Ψ_m for all $m \in \mathcal{C}(n)$ and all $n \in \overline{\mathcal{N}}$ (for instance, the latter is satisfied for the leaf nodes). In this case, the set $\text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ remains constant over the iterations, so we drop the iteration index for simplicity.

Let K with $|K| \in \mathbb{N}$ denote the finite set of facets of $\text{epi}(\overline{\text{co}}(\underline{Q}_n))$. We use the symbol F to refer to specific facets. Let $\overline{K} \subseteq K$ denote the subset of facets in which $(\bar{x}_{a(n)}, \bar{\theta}_n)$ is contained, *i.e.*, $(\bar{x}_{a(n)}, \bar{\theta}_n) \in F_k$ for all $k \in \overline{K}$. Analogously, let $\widehat{K}^\nu \subseteq K$ denote the subset of facets in which $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ is contained. Based on Lemma 4.3, we obtain the following result, which is proven in Appendix B.10.

Lemma 4.4. *There exists some $\widehat{\nu} \in \mathbb{N}$ such that for all $\nu \geq \widehat{\nu}$ we have $\widehat{K}^\nu \subseteq \overline{K}$.*

As $\widehat{K}^\nu \neq \emptyset$ by definition, this implies that for sufficiently large ν, the points $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ and $(\bar{x}_{a(n)}, \bar{\theta}_n)$ are located on a joint facet. Due to the convergence result in Lemma 4.3, this is intuitively clear. The case $\widehat{K}^\nu \not\subseteq \overline{K}$ is excluded by this convergence result. However, the case $\widehat{K}^\nu \subset \overline{K}$ is possible if $(\bar{x}_{a(n)}, \bar{\theta}_n)$ is located at the boundary of some facets.

We require some further auxiliary results.

Lemma 4.5. *There exists some $\bar{\nu} \in \mathbb{N}$ such that for all $\nu \geq \bar{\nu}$ the point $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ is equal to $(\bar{x}_{a(n)}, \bar{\theta}_n)$ or not a vertex of* $\text{epi}(\overline{\text{co}}(\underline{Q}_n))$.

This result follows immediately from Lemma 4.3. For sufficiently large ν, $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ may only be a vertex if $(\bar{x}_{a(n)}, \bar{\theta}_n)$ is a vertex and if they are equal.

Lemma 4.6. *Consider a convex polyhedron S. Let $x^1, x^2 \in S$ be two points located on a joint face (not necessarily a facet) $\mathfrak{F}$ of S with $x^2 \in \text{relint}(\mathfrak{F})$. Then, a cut supporting S at x^2 is also supporting S at x^1.*

We provide a proof in Appendix B.11. We are now able to prove our first main result, stating that for sufficiently large ν a deep Lagrangian cut supporting $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ will also support $(\bar{x}_{a(n)}, \bar{\theta}_n)$.

Lemma 4.7. *For any $\overline{\mathcal{N}}$, consider subproblem (6) with fixed approximations Ψ_m for all $m \in \mathcal{C}(n)$. Moreover, consider a subsequence of trial points $(\bar{x}_{a(n)}, \theta_n^{i_\nu})_{\nu\in\mathbb{N}}$ with fixed first component $\bar{x}_{a(n)} \in X_{a(n)}$ for all $\nu \in \mathbb{N}$. Then, there exists some $\check{\nu} \in \mathbb{N}$ such that for all $\nu \geq \check{\nu}$, a deep Lagrangian cut as defined in Definition 3.15 supports* $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n))$ *at* $(\bar{x}_{a(n)}, \bar{\theta}_n)$.

We provide a proof in Appendix B.12. We can now turn to properties of the cuts generated within NBD. Here, we exploit that $x_n \in \{0,1\}^{d_n}$ for all $n \in \mathcal{N}$.

Theorem 4.8. *Let $X_n = \{0,1\}^{d_n}$ for all $n \in \mathcal{N}$, and let Assumption 4 be satisfied. Then, after a finite number of iterations i, for all $n \in \overline{\mathcal{N}}$ the approximations Ψ_n^{i+1} are exact for* $\operatorname{epi}(Q_n(\cdot))$ *at all $x_n^i \in X_{a(n)}$ computed in the forward pass.*

A proof is presented in Appendix B.13.

4.3 Convergence Result

Based on the results in Theorem 4.2 or Theorem 4.8, respectively, finite convergence of NBD can be concluded. For more details, see [31].

Corollary 4.9. *Let $X_n = \{0,1\}^{d_n}$ for all $n \in \mathcal{N}$, and let Assumption 4 be satisfied. Assume that NBD is applied with generation of deep Lagrangian cuts or LN Lagrangian cuts satisfying Theorem 3.27 (ii). Then, NBD terminates with an optimal policy for (MS-MILP) after a finite number of iterations.*

We should emphasize again that this result only holds because of the finiteness and the binary character of X_n, so it is not necessarily true for general continuous state spaces. Furthermore, in the above proofs we use the idea that for any $x_{a(n)} \in X_{a(n)}$ after finitely many steps the deep or LN Lagrangian cuts will coincide with the original Lagrangian cuts, *i.e.* become *tight* in the sense of [31]. While this seems to imply more iterations than classical NBD (or SDDiP), the vision is that in practice convergence may actually be achieved faster.

5 Computational Experiments

We report results for a computational study of SDDiP using the proposed cut generation framework. For comparison, we also run tests using established cut generation techniques in SDDiP. More precisely, we consider the following approaches to generate cuts:

- `B`: Classical Benders cuts using either a single-cut or a multi-cut approach.
- `SB`: Strengthened Benders cuts [31] using either a single-cut or a multi-cut approach.
- `L`: Classical Lagrangian cuts from Appendix A using either a single-cut or a multi-cut approach.
- $\ell^1, \ell^\infty, \ell^{1\infty}$: Deep Lagrangian cuts from Definition 3.15 for the ℓ^1-norm, the ℓ^∞-norm and a linear combination of $0.5\|\cdot\|_1 + 0.5\|\cdot\|_\infty$.

- `LN`: LN Lagrangian cuts from Definition 3.22 using the `Mid`, `In-Out`, `Eps`, `Relint` and `Conv` heuristics from Sect. 3.6.2 to identify core points and to determine the normalization coefficients. For `Eps` we use a perturbation of the incumbent by 10^{-6}.

For now, we do not test the improvement techniques `epi-state` or `primal-obj` from Sect. 3.6.2. By construction, all deep and LN cuts require using a multi-cut approach.

We test the proposed methods on different instances of a capacitated lot-sizing problem (CLSP), which is described in the appendix of Trigeiro et al. [29] and has stagewise independent uncertain demand for each product. This problem is also considered in Ahmed et al. [1] and identified to be challenging for exact decomposition methods like SDDiP.

In our experiments, we consider instances with 3 or 10 state variables, 4, 6, 10 or 16 stages and 20 realizations of the uncertain demand at each stage. For the case of 3 state variables and 20 realizations per stage, we use the exact same scenarios as in Ahmed et al. [1], for larger instances we use the same methodology to generate scenarios.

For instances with 3 state variables we apply SDDiP with a binary approximation of the continuous state variables with discretization precision of 1.0 [see 31]. This means that the modified MS-MILP which is tackled by SDDiP contains 30 state variables. For instances with 10 state variables initially, using a binary approximation would produce state dimensions which are computationally intractable for SDDiP, as its complexity grows exponentially in the dimension of the state space [30]. Therefore, in these cases, we apply SDDiP without a binary approximation.

We describe our implementation of SDDiP and our parameter choices in more detail in Sect. 5.1. Then, we discuss results for experiments of CLSP with state binarization, where we compare the above cut generation techniques when they are applied individually (Sect. 5.2) as well as combined with `SB` cuts (Sect. 5.3). In Sect. 5.4 we study a restriction of the dual space. In Sect. 5.5 we deal with larger instances of CLSP where no binary approximation is applied.

5.1 Implementation Details

SDDiP and all cut generation approaches are implemented in Julia-1.5.3 [5] based on the existing packages `SDDP.jl` [15] and `JuMP.jl` [16]. The code is available on GitHub as part of a larger project called `DynamicSDDiP.jl` (see `https://github.com/ChrisFuelOR/DynamicSDDiP.jl`).

SDDiP is terminated after a predefined time limit or if the obtained lower bounds start to stall. In each forward pass, one scenario path is randomly sampled. After termination of SDDiP, an in-sample Monte Carlo simulation with 1000 replications is conducted on the finite scenario tree to compute a statistical upper bound for the current policy.

The Lagrangian dual problems in SDDiP are solved using a level bundle method with a maximum of 1000 iterations and an optimality tolerance of 10^{-4}. The multipliers π_n are initialized with a vector of zeros and π_{n0} is initialized with 1. Sometimes the level bundle method reports infeasibilities in the quadratic auxiliary problem. In that case, we proceed with a standard Kelley step instead. Moreover, in the case of different numerical issues in solving the Lagrangian dual, the solution process is stopped and a valid cut is constructed with the current values of the multipliers.

For the LN cuts, as pointed out before, the choice of normalization coefficients (u_n, u_{n0}) is crucial for the cut quality, but also to achieve a bounded subproblem. If the

chosen heuristic yields *only* coefficients (close to or) equal to zero, in our implementation no cut is generated at all. Moreover, non-zero coefficients may not yield a bounded dual problem if they correspond to a direction $(u_n, u_{n0}) \notin \operatorname{cone}\left(\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1})) - (x^i_{a(n)}, \theta^i_n)\right) \setminus \{0\}$, see Sect. 3.6.2. We use infeasibility of problem (20) as a flag for possible unboundedness. In that case, we add some artificial bounds to the dual problem (14) to at least generate some valid cuts. Some pre-testing indicated that for `LN-Mid`, `LN-Relint` and `LN-Conv` using multipliers bounds $\pi_n \in [-10, 10]^{d_{a(n)}}, \pi_{n0} \in [0, 10]$ and for `LN-Eps` and `LN-In-Out` introducing an artificial objective bound of 1000 leads to reasonable results. Additionally, note that if problem (20) is feasible, we obtain an upper bound for the optimal value of the dual problem (14). We can use this bound to ensure boundedness for the approximating models in the level bundle method.

All occuring LP, MILP and QP subproblems are solved using Gurobi 9.0.3 with an optimality tolerance of 10^{-4} and a time limit of 300 seconds. All tests with binary approximation are run on a Windows machine with 64 GB RAM and an Intel Core i7-7700K processor (4.2 GHz). All tests without binary approximation are run on a Windows machine with 128 GB RAM and an Intel Xeon E5-1630v4 processor (3.7 GHz).

5.2 Comparison with Classical Lagrangian and Benders Cuts

For our first experiments we apply SDDiP (with binary approximation) to CLSP instances with 3 state variables. We only use one type of cut for the whole solution process. The results are illustrated in several figures throughout this section. The full results are provided in Appendix D.

First, we consider experiments with $T = 4$ or $T = 6$ stages and a maximum run time of 3 hours and 4 hours, respectively. The obtained lower bounds are depicted in Fig. 11. We observe that `B` and `SB` do not manage to close the optimality gap and that the obtained lower bounds stall very fast. `L`, whereas better in theory, leads to even worse lower bounds. One reason is that solving the dual problems is computationally costly, but additionally, compared to `SB` the tighter cuts seem to lead to worse incumbents on earlier stages or in following iterations. In fact, even in the first iteration, the lower bound obtained by `SB` is superior to that obtained by `L`. Many variants of deep and LN Lagrangian cuts outperform `SB` and `L` with respect to the lower bounds and gaps, even if the optimality gap is not completely closed in the predefined time horizon. While the quality of the lower bounds is better, the iteration times and the number of iterations in the bundle method are not necessarily reduced despite solving a bounded problem, especially not for `LN` cuts.

Among the new approaches, ℓ^∞-deep cuts perform rather bad. Our hypothesis is that for binary state variables this approach is very prone to degeneracy in the normalized Lagrangian dual problems, which then leads to cuts of bad quality, see Sect. 1. To test this hypothesis, we perform experiments using an additional optimization step that was proposed in an earlier version of `SDDP.jl` and resembles the two-step cut generation in [22]. In this step we minimize the norm among all optimal solutions of the Lagrangian dual. Therefore, we label this approach as `MNC` (minimal norm choice). With `MNC`, the bounds obtained by ℓ^∞-deep cuts improve considerably. For this reason, without further notice, we always use this approach in the following experiments. For all other cut generation approaches, we observe no significant improvement in lower bounds per time using an `MNC` step.

We now consider experiments with more stages, $T = 10$ and $T = 16$ to be precise,

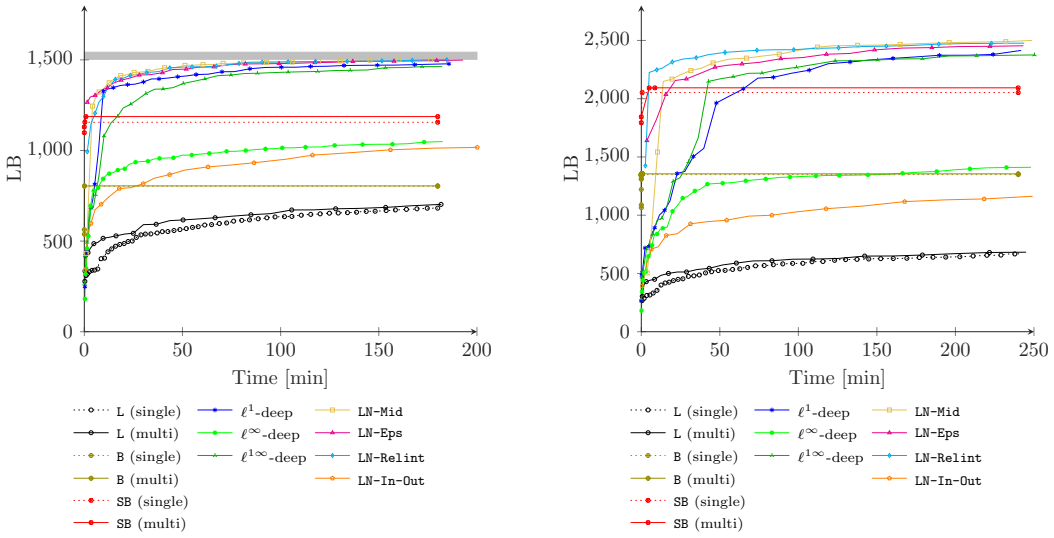


Figure 11: Lower bound development over time for experiments on CLSP with state binarization.

Note. LEFT: $T = 4$. RIGHT: $T = 6$. For $T = 4$, the shaded gray area is where the optimal value lies according to an approximate solution of the deterministic equivalent. For `B` and `SB`, SDDiP quickly terminates due to stalling lower bounds, so the last lower bound is interpolated over the whole time horizon. Marks at every second iteration, except for `B` and `SB`.

with run times of 5 and 8 hours, respectively. The obtained lower bounds are depicted in Fig. 12. We observe that deep cuts perform mediocre for 16 stages. `LN` cuts perform best with respect to the lower bounds, but for 16 stages hardly outperform `SB`. This is mainly due to long iteration times, even in comparison to deep cuts, see Fig. 13. Moreover, as Fig. 14 shows, an improvement in lower bounds does not necessarily translate to an improvement of the obtained policies. In fact, `SB` achieves the best simulated upper bounds. It seems that using `SB` it is possible to quickly identify good feasible solutions, but that the lower bounds are too loose to get a certificate for optimality, whereas for Lagrangian cuts it is the opposite.

5.3 Combination with SB Cuts

As shown in the previous section, using SDDiP with *only* Lagrangian cuts becomes extremely slow for large problems. Therefore, in practice, it is reasonable to combine different types of cuts. Already in the original SDDiP work [31] it is proposed to combine Lagrangian cuts, which can provide convergence guarantees, and strengthened Benders cuts, which can be computed efficiently.

To evaluate the performance of deep and LN cuts in this setting, we conduct experiments where we start with only `SB` for the first 20 iterations to get a quick bound improvement, and then generate SB cuts and Lagrangian cuts in each iteration. The lower bounds are depicted in Fig. 15, while the simulation results and optimality gaps are presented in Fig. 14 next to the ones of the previous case.

The lower bound results heavily improve for `L`, but are not affected too much for deep or `LN` cuts. They are still better for these approaches than the ones obtained using only `SB`, though. As the simulated upper bounds are better than in the previous setting for all types of cuts, we can conclude that a combination of SB cuts and Lagrangian

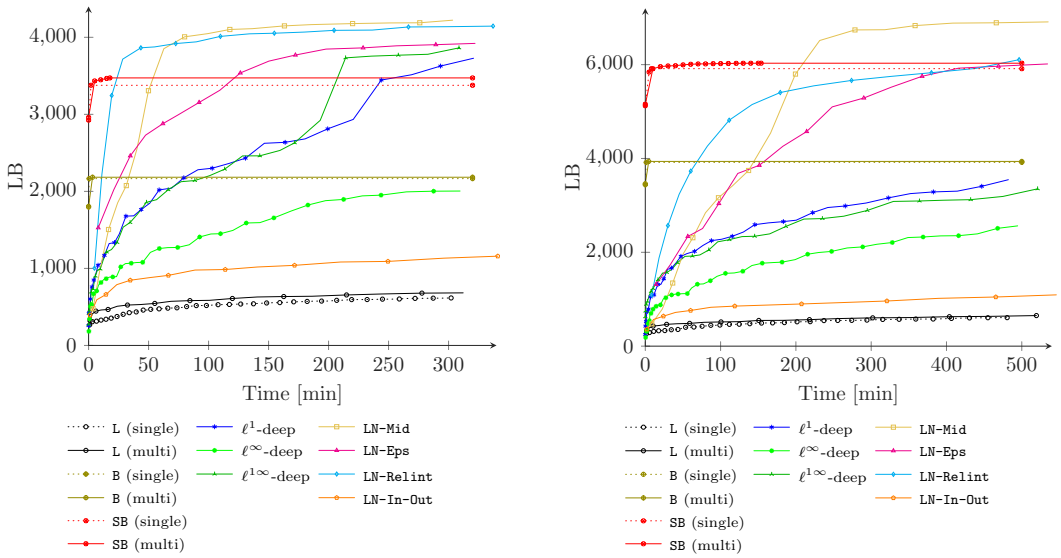


Figure 12: Lower bound development over time for experiments on CLSP with state binarization.

Note. LEFT: $T = 10$. RIGHT: $T = 16$.

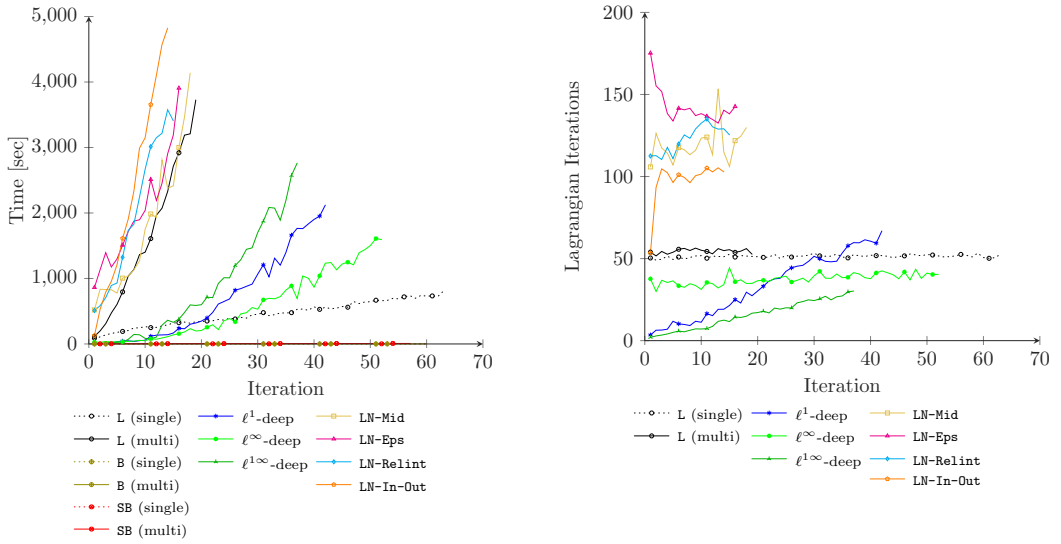


Figure 13: Different analyses for experiments on CLSP with state binarization and $T = 16$.

Note. LEFT: Time per iteration of SDDiP. RIGHT: Iterations required to solve Lagrangian dual over iterations of SDDiP.

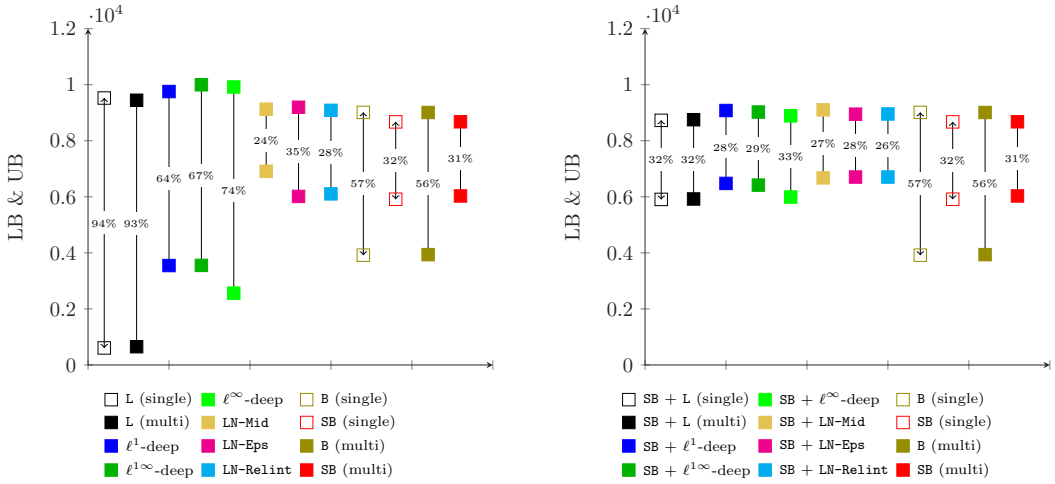


Figure 14: Optimality gaps for experiments on CLSP with state binarization with $T = 16$.

Note. LEFT: Run with only one type of cuts. RIGHT: Runs with SB plus additional cuts from iteration 21.

cuts combines the advantages of good lower bounds and reasonably good simulation results for the policies.

5.4 The Chen-Luedtke Approach: Restricting the Dual Space

Another approach suited to accelerate SDDiP was recently put forward by Chen and Luedtke [11] for the two-stage case. They propose to restrict the feasible set of the normalized Lagrangian dual problem (14) to a small subset of valid multipliers (π_n, π_{n0}). More precisely, the idea is to restrict the multipliers π_n to the span of a set of previously generated Benders cut coefficients $\widehat{\pi}_n^k, k = 1, \ldots, K$, for some predefined parameter K. That is, we introduce the constraint

$$\pi_n = \sum_{k=1}^{K} \gamma_{nk} \widehat{\pi}_n^k,$$

which also means that we add variables $\gamma_{nk}, k = 1, \ldots, K$, to the dual problem. While we lose tightness and convergence guarantees using this dual space restriction, the search space for the level Bundle method is significantly reduced, so that cuts can be generated faster. We refer to this as the CL approach.

In principle, the CL approach can be combined with any of the previously used normalization techniques. Additionally, it allows for an alternative normalization [11]. Instead of using $g(\pi_n, \pi_{n0}) = \|\pi_n, \pi_{n0}\|_1$, we may as well use some normalization function $g(\pi_n, \pi_{n0}, \gamma_n) = \|\gamma_n, \pi_{n0}\|_1$. This choice should lead to solutions with sparse γ. In the following, we denote this approach by CL-γ.

For our computational experiments, we choose $K = 20$. Apart from the dual space restriction, we use the same setting as for the previous runs (20 iterations of only SB, after that SB and Lagrangian cuts).

The results are depicted in Fig. 16. The number of Lagrangian iterations and the time per iteration are reduced significantly. Moreover, compared to only using La-

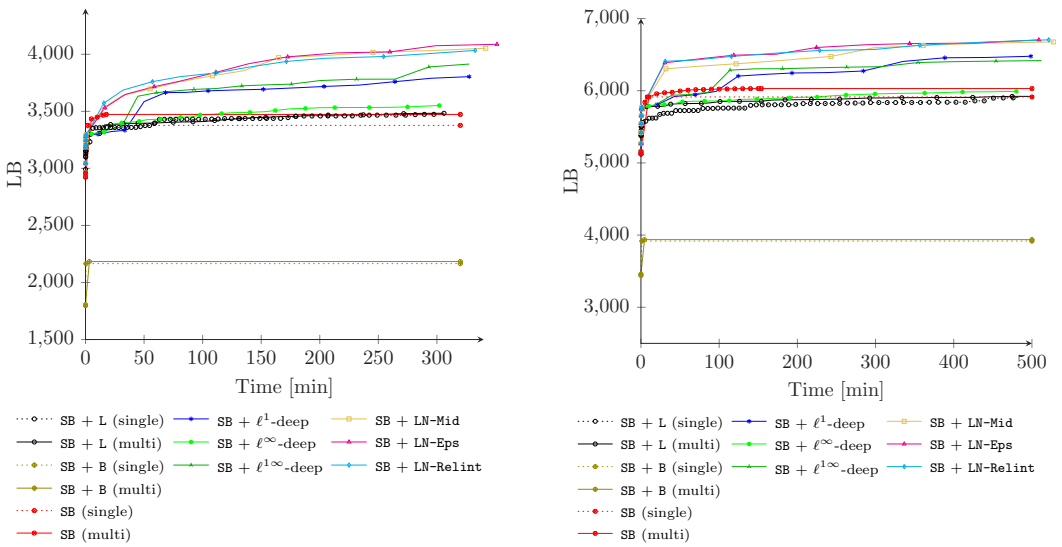


Figure 15: Lower bound development over time for experiments on CLSP with state binarization using a combination with SB cuts.

Note. LEFT: $T = 10$. RIGHT: $T = 16$. 20 iterations with SB and then SB and Lagrangian cuts in each iteration.

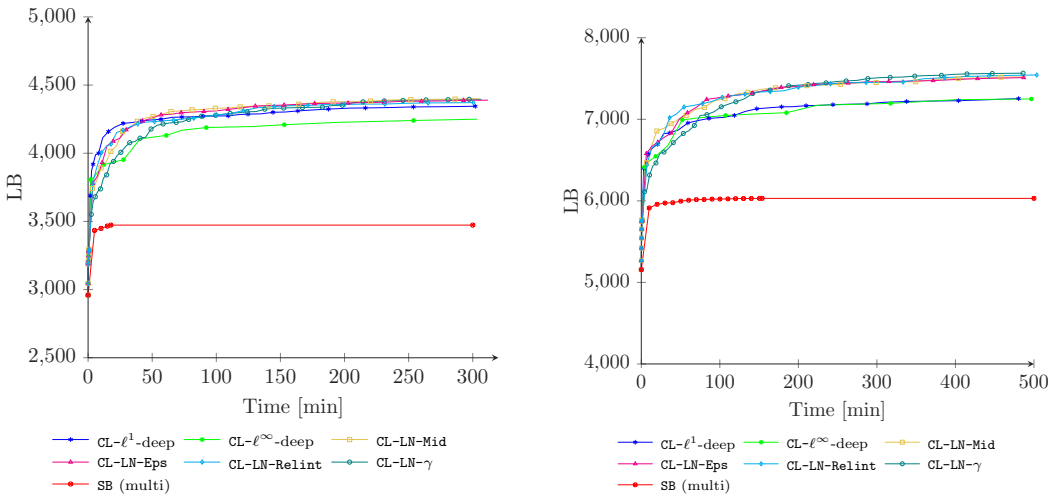


Figure 16: Lower bound development over time for experiments on CLSP with state binarization using the Chen-Luedtke approach.

Note. LEFT: $T = 10$. RIGHT: $T = 16$. 20 iterations with SB and then SB and Lagrangian cut in each iteration with dual space restriction for $K = 20$.

grangian cuts, much better lower bounds are obtained in the same time. This implies that the lower bounds for the combined approach are also much better than for only using `SB`. Interestingly, the lower bounds are even better *per iteration* than without dual space restriction, similar to what we observed for `SB` before. This illustrates that the quality of cuts does not only depend on tightness or depth, but also on the incumbents which they induce in the previous stages and following iterations. The chosen normalization approach seems not to be decisive in this setting.

5.5 Results for Larger Instances

As discussed before, for experiments with a larger state dimension, we do not apply a binary approximation. This means that we do not have convergence guarantees and cannot expect the optimality gap to be closed. However, in return iterations should take considerably less time. Note that this is often the go-to approach to apply SDDP-like methods to mixed-integer programs in practice.

In this case, we also consider `LN-Conv` with convex combination parameters of 0.5, 0.75, 0.9 and 0.99 (with higher values encoding a higher proximity to $\left(x^i_{a(n)}, \underline{Q}^{i+1}_n(x^i_{a(n)})\right)$).

The results show that deep and `LN` Lagrangian cuts manage to achieve better lower bounds and gaps than conventional cuts, see Fig. 17 and 18. Using a dual space restriction allows to further reduce the optimality gap to about 21% in 3 hours and to about 19% in 5 hours. This is a considerable improvement compared to Benders cuts, which are most frequently used in practice. This means that the proposed cut generation techniques may be helpful to improve the convergence behavior even if no binary approximation is applied. However, as for the previous experiments, we cannot conclude that the improvement in lower bounds necessarily leads to an improvement of the in-sample performance of the obtained policy, and thus to better optimality gaps.

We also observe that the average iteration time is reduced considerably compared to the previous test cases with binary state approximation (about 60-75% reduction). For this reason, and because the state space is also lower-dimensional, even without convergence guarantees, better optimality gaps are obtained than for the previous test cases in the same time.

5.6 Discussion and Potential Improvements

Overall, our results show significant improvements of the obtained lower bounds using the new cut generation framework in all cases: with state binarization, without state binarization, combined with strengthened Benders cuts or applying the cuts on their own. With binarization, especially LN cuts yield strong improvements, whereas without binarization also deep cuts perform reasonably well. We see that better lower bounds do not necessarily translate to better performances of the obtained policies, though. Additionally, we observe that even using the new framework, SDDiP suffers from well-known computational drawbacks such as high computational cost to solve Lagrangian dual problems (especially if a binary approximation of the state space is applied) and slow convergence of lower bounds due to premature stalling [2], so even after hours of run time the observed optimality gaps are still considerable.

In the future, the performance of SDDiP including our proposed cut generation framework could be improved in several ways. First, the solution of independent Lagrangian duals for nodes $m \in \mathcal{C}(n)$ could be parallelized. Second, potential warm starting or acceleration techniques for the Lagrangian dual (*e.g.* using sub-optimal so-

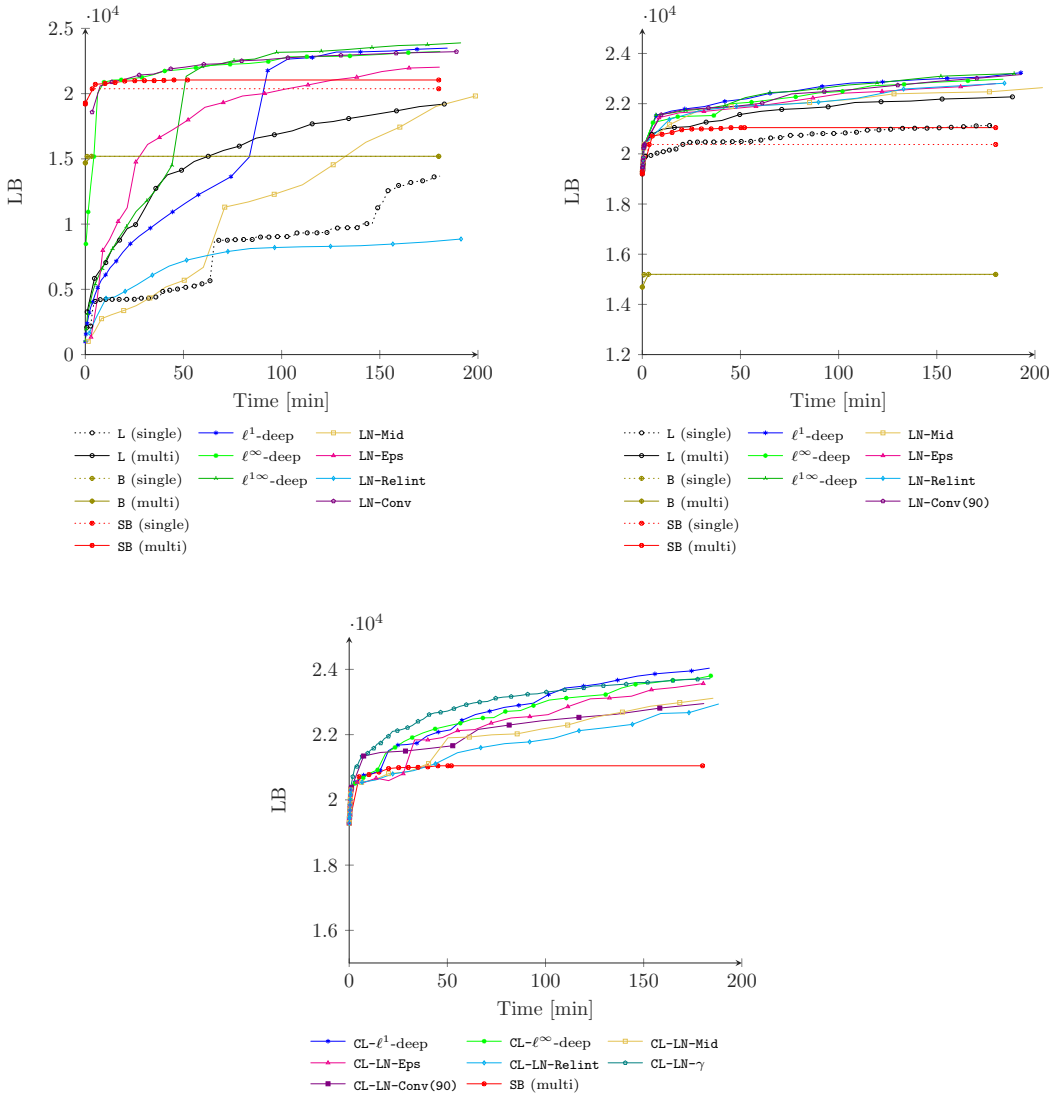


Figure 17: Lower bound development over time for experiments on CLSP with $T = 16$ and without state binarization.

Note. UP LEFT: Run with only one type of cuts. UP RIGHT: Runs with `SB` plus additional cuts from iteration 21. BELOW: Runs with Chen-Luedtke approach for $K = 20$.

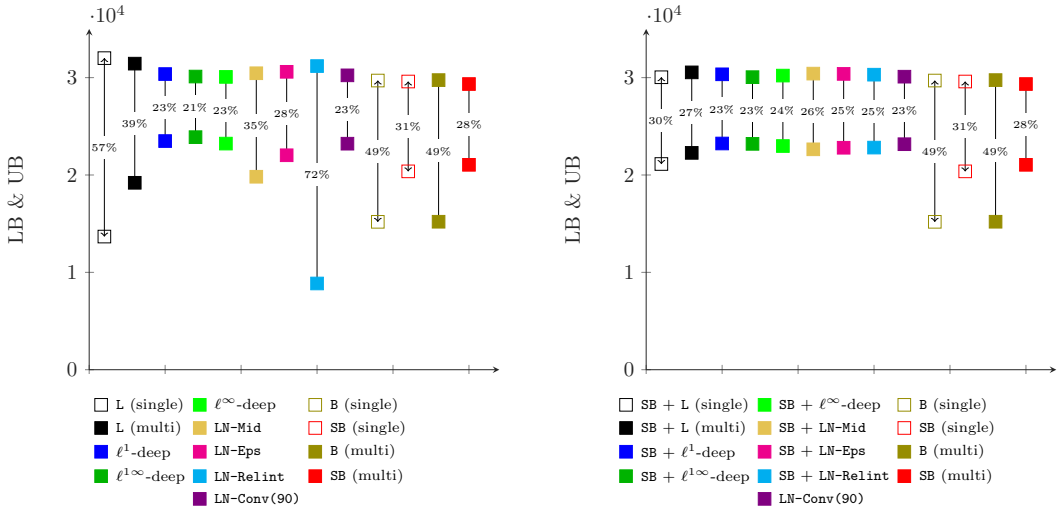


Figure 18: Optimality gaps for experiments on CLSP with $T = 16$ and without state binarization.

Note. LEFT: Run with only one type of cuts. RIGHT: Runs with `SB` plus additional cuts from iteration 21.

lutions) could be explored, even if challenging for multistage problems. Third, the dual space restriction suggested by Chen and Luedtke [11] looks promising to reduce the computational effort while not compromising cut quality by too much. We think that future research could focus more on priorly restricting the dual space to reduce the computational effort for solving Lagrangian dual problems.

Our computational experiments also reveal that, especially for LN cuts, SDDiP may occasionally suffer from numerical issues, and that the obtained results show a high sensitivity with respect to the chosen parameters. First of all, it is a common issue of cutting-plane methods that they may lead to ill-conditioned problems if the cut and problem coefficients are not properly scaled. In this context, an appropriate choice of normalization coefficients (u_n, u_{n0}) for LN cuts is crucial. In addition, core point identification in general remains a challenging task, especially when integer requirements are apparent. Addressing these challenges in detail merits further research.

Finally, we have only considered a very specific test problem so far. For a more profound and general performance assessment, therefore experiments for more and also larger test problems have to be carried out. In fact, we are planning to enhance our experiments with tests of a capacited facility location problem with pure binary state variables and local integer constraints. This will also allow us to further explore the challenges of core point identification in the context of integer requirements.

6 Conclusion

In this article, we propose a new framework to generate Lagrangian cuts for value functions occurring in MS-MILPs, which generalizes earlier proposals for 2-stage problems. We prove that using different normalizations of the Lagrangian dual problems, cuts with different favorable properties can be obtained, such as maximal depth, being facet-defining or Pareto-optimal. Our framework allows for a lot of flexibility in cut generation, and thus notably extends the toolbox of SDDiP. If all state variables are

binary, finite almost sure convergence of SDDiP is assured, as for classical Lagrangian cuts.

We provide computational results for experiments on a capacitated lot-sizing problem. The results show that the lower bounds in SDDiP can be vastly improved by incorporating our proposed framework, although not eliminating other well-known computational drawbacks, such as excessive computational effort, slow convergence and inability to close the optimality gap. As described in the previous section, therefore more theoretical and computational research is required to efficiently apply our proposed framework, and SDDiP in general, on large-scale problems in practice.

Finally, our cut generation framework requires a multi-cut approach, whereas for multistage problems often a single-cut approach is favored computationally, as much less cuts have to be added per iteration. Trying to compute deep Lagrangian cuts or LN Lagrangian cuts in a single-cut framework in a computationally efficient way could therefore be an interesting research direction.

Acknowledgements

Andy Sun's research is partially funded by the National Science Foundation CAREER award 2316675. Christian Füllner's research is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–445857709. This work was started during Christian Füllner's research visit at Georgia Institute of Technology, which was funded by the Karlsruhe House of Young Scientists (KHYS). The authors thank Filipe Cabral for providing data for the capacited lot-sizing problem considered in the computational experiments.

A Classical Lagrangian Cuts

For any $n \in \overline{\mathcal{N}}$, let $\mathfrak{Q}_n^{i+1}(\cdot)$ denote the current cut approximation for value function $Q_n(\cdot)$ in a decomposition method, such as SDDiP. Then, a Lagrangian cut can be generated by considering a special Lagrangian relaxation of the nodal subproblem (that is, subproblem (3) with $Q_m(\cdot), m \in \mathcal{C}(n)$, replaced by $\mathfrak{Q}_m^{i+1}(\cdot)$). More precisely, the copy constraints $z_n = x_{a(n)}^i$ are relaxed using a given vector of dual multipliers $\pi_n \in \mathbb{R}^{d_{a(n)}}$, which yields

$$\mathcal{L}_n^{i+1}(\pi_n) := \min_{x_n, y_n, z_n, (\theta_m)} \Big\{ f_n(x_n, y_n) + \sum_{m \in \mathcal{C}(n)} p_{nm}\theta_m - \pi_n^\top z_n \ : \ (z_n, x_n, y_n) \in \mathcal{F}_n, \\ z_n \in Z_{a(n)}, \theta_m \geq \mathfrak{Q}_m^{i+1}(\cdot)(x_n), \ m \in \mathcal{C}(n) \Big\}.$$

For varying π_n, this relaxation defines the *dual function* $\mathcal{L}_n^{i+1}(\cdot)$. The problem of optimizing the dual function over the dual multipliers π_n is the *Lagrangian dual problem*:

$$\max_{\pi_n} \Big\{ \mathcal{L}_n^{i+1}(\pi_n) + \pi_n^\top x_{a(n)}^i \Big\}. \tag{21}$$

By solving problem (21), a Lagrangian cut for $Q_n(\cdot)$ can be derived as

$$\theta_n \geq \mathcal{L}_n^{i+1}(\pi_n^i) + (\pi_n^i)^\top x_{a(n)}, \tag{22}$$

where π_n^i denotes feasible dual multipliers in (21) for node n [31]. If required, feasibility cuts can be derived in a similar fashion [see 11, 25].

The Lagrangian cuts (22) have useful properties [31]. Their right-hand sides are valid under-estimators of $Q_n(\cdot)$ and tight at $\overline{\text{co}}(\underline{Q}_n^{i+1})(x_{a(n)}^i)$ (given that optimal dual multipliers π_n^i are used). Moreover, only finitely many different Lagrangian cuts exist if only dual basic solutions are considered. Finally, if the state variables x_n are binary, the cuts are even tight at $\underline{Q}_n^{i+1}(x_{a(n)}^i)$. These properties ensure almost sure finite convergence of SDDiP.

B Proofs

In this section, we present the proofs that are not displayed in the main text.

B.1 Proof of Lemma 3.7

Proof. This is proven in [11] for $\text{epi}(Q_n)$, but we provide a customized proof here. Let $(x_{a(n)}, \theta_n) \in \text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$. Then,

$$\begin{aligned}
(\pi_n^i)^\top x_{a(n)} + \pi_{n0}^i \theta_n &\geq \min_{x_{a(n)}, \theta_n} \Big\{ (\pi_n^i)^\top x_{a(n)} + \pi_{n0}^i \theta_n \ : \ (x_{a(n)}, \theta_n) \in \text{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1})) \Big\} \\
&= \min_{\lambda_n, z_n} \Big\{ (\pi_n^i)^\top z_n + \pi_{n0}^i c_n^\top \lambda_n \ : \ (\lambda_n, z_n) \in \text{conv}(\mathcal{W}_n^{i+1}) \Big\} \\
&= \min_{\lambda_n, z_n} \Big\{ (\pi_n^i)^\top z_n + \pi_{n0}^i c_n^\top \lambda_n \ : \ (\lambda_n, z_n) \in \mathcal{W}_n^{i+1} \Big\} \\
&= \mathscr{L}_n^{i+1}(\pi_n^i, \pi_{n0}^i).
\end{aligned}$$

The inequality follows by feasibility. The first equality uses the same relation that is also applied in (11). The second equality exploits that the objective function is linear, and the last one follows from the definition of $\mathscr{L}_n^{i+1}(\cdot)$ in (9). The second part of the assertion follows with $\operatorname{epi}(Q_n) \subseteq \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$. □

B.2 Proof of Lemma 3.10

Proof. According to Brandenberg and Stursberg [9], $\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i)$ can be rewritten as

$$\begin{aligned}\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i) = \Big\{ &(\gamma_n, \gamma_{n0}) \in \mathbb{R}^{d_{a(n)}} \times \mathbb{R} \ : \\ &\gamma_n^\top x_{a(n)}^i + \gamma_{n0}\theta_n^i - \operatorname{supp}_{\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))}(\gamma_n, \gamma_{n0}) \geq 1 \Big\},\end{aligned} \tag{23}$$

where $\operatorname{supp}_{\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))}(\cdot)$ denotes the support function of $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$. This function can be expressed as follows:

$$\begin{aligned}&\operatorname{supp}_{\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))}(\gamma_n, \gamma_{n0}) \\ &= \max_{x_{a(n)}, \theta_n} \left\{ \gamma_n^\top x_{a(n)} + \gamma_{n0}\theta_n \ : \ (x_{a(n)}, \theta_n) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1})) \right\} \\ &= \max_{x_{a(n)}, \theta_n, \lambda_n, z_n} \Big\{ \gamma_n^\top x_{a(n)} + \gamma_{n0}\theta_n \ : \ \tilde{A}_n\lambda_n + \tilde{B}_n z_n \geq \tilde{d}_n, z_n = x_{a(n)}, \theta_n - c_n^\top \lambda_n \geq 0 \Big\} \\ &= \min_{\mu_n, \pi_n, \pi_{n0}} \Big\{ \tilde{d}_n^\top \mu_n \ : \ \tilde{A}_n^\top \mu_n - c_n \pi_{n0} = 0, \tilde{B}_n^\top \mu_n - \pi_n = 0, \\ &\qquad\qquad \pi_n = \gamma_n, \pi_{n0} = \gamma_{n0}, \pi_{n0} \leq 0, \mu_n \leq 0 \Big\} \\ &= \min_{\mu_n} \Big\{ -\tilde{d}_n^\top \mu_n \ : \ -\tilde{A}_n^\top \mu_n - c_n \gamma_{n0} = 0, -\tilde{B}_n^\top \mu_n - \gamma_n = 0, \gamma_{n0} \leq 0, \mu_n \geq 0 \Big\}.\end{aligned} \tag{24}$$

The first equation applies the definition of support functions. The second one follows from Remark 3.3 and the third one exploits strong duality for LPs. We insert (24) into (23), and observe that the set remains unchanged if we replace the minimum operator using an existence quantor. □

B.3 Proof of Lemma 3.14

Proof. According to Theorem 3.5, we have $\widehat{v}_n^{D,i+1}(x_{a(n)}^i, \theta_n^i) = 0$ for the non-normalized Lagrangian dual (10). By definition of (10) and its normalization (14), we can thus conclude $\widehat{v}_n^{ND,i+1}(x_{a(n)}^i, \theta_n^i) \leq \widehat{v}_n^{D,i+1}(x_{a(n)}^i, \theta_n^i) = 0$.

Let $(\widehat{\pi}_n, \widehat{\pi}_{n0})$ be an optimal point for problem (10), *i.e.*, $\mathscr{L}_n^{i+1}(\widehat{\pi}_n, \widehat{\pi}_{n0}) - (\widehat{\pi}_n)^\top x_{a(n)}^i - \widehat{\pi}_{n0}\theta_n^i = 0$. If $\|\widehat{\pi}_n, \widehat{\pi}_{n0}\| \leq 1$, then it is also feasible for (14). As the objective of both problems is the same, $\widehat{v}_n^{ND,i+1}(x_{a(n)}^i, \theta_n^i) = 0$.

Otherwise, there exists $\mu > 0$ such that $\frac{1}{\mu}(\widehat{\pi}_n, \widehat{\pi}_{n0})$ is feasible for (14). By feasibility, it follows

$$\begin{aligned}\widehat{v}_n^{ND,i+1}(x_{a(n)}^i, \theta_n^i) &\geq \mathscr{L}_n^{i+1}\left(\frac{1}{\mu}\widehat{\pi}_n, \frac{1}{\mu}\widehat{\pi}_{n0}\right) - \frac{1}{\mu}(\widehat{\pi}_n)^\top x_{a(n)}^i - \frac{1}{\mu}\widehat{\pi}_{n0}\theta_n^i \\ &= \frac{1}{\mu}\Big(\mathscr{L}_n^{i+1}(\widehat{\pi}_n, \widehat{\pi}_{n0}) - (\widehat{\pi}_n)^\top x_{a(n)}^i - \widehat{\pi}_{n0}\theta_n^i\Big) = 0,\end{aligned} \tag{25}$$

where we exploited that $\mathscr{L}_n^{i+1}(\cdot)$ is positive homogeneous. The reverse direction can be

shown in a similar way. □

B.4 Proof of Lemma 3.23

Proof. By definition, the normalized Lagrangian dual problem is equivalent to

$$\begin{aligned}&\max_{(\pi_n,\pi_{n0})\in\Pi_n}\left\{\min_{(\lambda_n,z_n)\in\mathcal{W}_n^{i+1}}\left\{\pi_n^\top(z_n-x^i_{a(n)})+\pi_{n0}(c_n^\top\lambda_n-\theta_n^i)\right\}\right\}\\ &=\max_{(\pi_n,\pi_{n0})\in\Pi_n}\left\{\min_{(\lambda_n,z_n)\in\mathrm{conv}(\mathcal{W}_n^{i+1})}\left\{\pi_n^\top(z_n-x^i_{a(n)})+\pi_{n0}(c_n^\top\lambda_n-\theta_n^i)\right\}\right\}\end{aligned}$$

with $\Pi_n := \left\{(\pi_n,\pi_{n0})\in\mathbb{R}^{d_{a(n)}}\times\mathbb{R} \;:\; \pi_{n0}\geq 0, u_n^\top\pi_n+u_{n0}\pi_{n0}\leq 1\right\}$. The equation follows from linearity.

Using Remark 3.3 and then LP duality for the inner minimization problem, we obtain the equivalent problem

$$\begin{aligned}\max_{\pi_n,\pi_{n0},\mu_n}\Big\{\tilde{d}_n^\top\mu_n-\pi_n^\top x^i_{a(n)}-\pi_{n0}\theta_n^i \;:\; &\mu_n\geq 0,\ \pi_{n0}\geq 0,\ u_n^\top\pi_n+u_{n0}\pi_{n0}\leq 1,\\ &\tilde{A}_n^\top\mu_n-\pi_{n0}c_n=0,\ \tilde{B}_n^\top\mu_n-\pi_n=0\Big\}.\end{aligned}$$

This is an LP. Using LP duality and Remark 3.3 again, the assertion follows. □

B.5 Proof of Theorem 3.25

To prove this, we first require the definition and some results for the alternative polyhedron.

Definition B.1 ([17])**.** *The* alternative polyhedron *of* (10) *is defined as*

$$\mathcal{A}_n(x^i_{a(n)},\theta_n^i) := \left\{(\mu_n,\pi_n,\pi_{n0})\in\mathbb{R}^k\times\mathbb{R}^{d_{a(n)}}\times\mathbb{R} \;:\; \begin{array}{l}\mu_n,\pi_{n0}\geq 0\\ \tilde{A}_n^\top\mu_n-\pi_{n0}c_n=0\\ \tilde{B}_n^\top\mu_n-\pi_n=0\\ \tilde{d}_n^\top\mu_n-\pi_n^\top x^i_{a(n)}-\pi_{n0}\theta_n^i=1\end{array}\right\}.$$

If we relax the last equality constraint to $\tilde{d}_n^\top\mu_n-\pi_n^\top x^i_{a(n)}-\pi_{n0}\theta_n^i\geq 1$, we call the obtained set the relaxed alternative polyhedron *and denote it by $\widehat{\mathcal{A}}_n(x^i_{a(n)},\theta_n^i)$.*

As shown in [9], the sets $\widehat{\mathcal{A}}_n(x^i_{a(n)},\theta_n^i)$ and $\mathcal{R}_n^{i+1}(x^i_{a(n)},\theta_n^i)$ are closely related by a linear transformation, which in our case involves a unit matrix $I_{d_{a(n)}}$ of dimension $d_{a(n)}$.

Lemma B.2 (Theorem 2.1 in [9])**.** *The reverse polar set $\mathcal{R}_n^{i+1}(x^i_{a(n)},\theta_n^i)$ and the relaxed alternative polyhedron $\widehat{\mathcal{A}}_n(x^i_{a(n)},\theta_n^i)$ satisfy the relation*

$$\begin{pmatrix}\begin{pmatrix}0 & -I_{d_{a(n)}}\end{pmatrix} & 0\\ 0 & -1\end{pmatrix}\widehat{\mathcal{A}}_n(x^i_{a(n)},\theta_n^i)=\mathcal{R}_n^{i+1}(x^i_{a(n)},\theta_n^i).$$

Therefore, for problem (18) we obtain the related problem

$$\max_{\mu_n,\pi_n,\pi_{n0}}\left\{-u_n^\top\pi_n-u_{n0}\pi_{n0} \;:\; (\mu_n,\pi_n,\pi_{n0})\in\widehat{\mathcal{A}}_n^{i+1}(x^i_{a(n)},\theta_n^i)\right\}. \tag{26}$$

We can finally prove Theorem 3.25.

Proof. We first prove (i). It can be shown that the normalization constraint is binding at an optimal solution of the normalized Lagrangian dual problem (14), see the reasoning in Proposition 9 in [21]. Then, by using the reformulation ideas from Remark 3.3 and applying Theorem 3.20 from [28], this problem is equivalent to problem (26) in the following way:

Let $(\mu_n^*, \pi_n^*, \pi_{n0}^*)$ be optimal for (14) with optimal value $v_n^* > 0$. Then $(\bar{\mu}_n, \bar{\pi}_n, \bar{\pi}_{n0}) = \frac{1}{v_n}(\mu_n^*, \pi_n^*, \pi_{n0}^*)$ is optimal for (26) with optimal value $\bar{v}_n = -\frac{1}{v_n} < 0$. In reverse, let $(\bar{\mu}_n, \bar{\pi}_n, \bar{\pi}_{n0})$ be optimal for (26) with optimal value $\bar{v}_n < 0$. Then $(\mu_n^*, \pi_n^*, \pi_{n0}^*) = -\frac{1}{v_n}(\bar{\mu}_n, \bar{\pi}_n, \bar{\pi}_{n0})$ is optimal for (14) with optimal value $v_n^* = -\frac{1}{\bar{v}_n} > 0$. Note that the optimal values multiply to -1.

According to Corollary 2.2 in [9], which is based on the relation stated in Lemma B.2, $(\widehat{\gamma}_n, \widehat{\gamma}_{n0})$ is optimal for problem (18) if and only if there exists some $\widehat{\mu}_n$ such that $(\bar{\mu}_n, \bar{\pi}_n, \bar{\pi}_{n0}) = (\widehat{\mu}_n, -\widehat{\gamma}_n, -\widehat{\gamma}_{n0})$ is optimal for problem (26). Moreover, the optimal values $\widehat{v}_n$ and $\bar{v}_n$ are the same.

Hence, by the first step, we have a scaling relation between the optimal points of (14) and (26), and by the second step, we have a sign change relation between the optimal points of (26) and (18). The optimal values multiply to -1 after the first step and do not change in the second step. This proves the assertion.

We now prove (ii). Let $(\widehat{\gamma}_n, \widehat{\gamma}_{n0})$ be optimal for problem (18) with a valid certificate $\widehat{\mu}_n \geq 0$ in $\mathcal{R}_n^{i+1}(x_{a(n)}^i, \theta_n^i)$. Then, the valid cut

$$\tilde{d}_n^\top \widehat{\mu}_n + (\widehat{\gamma}_n)^\top x_{a(n)} + \widehat{\gamma}_{n0}\theta_n \leq 0$$

is induced, separating $(x_{a(n)}^i, \theta_n^i)$ from $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$ [13]. By the optimality relations used to prove (i) and by multiplying with $-\frac{1}{\widehat{v}_n} = -\frac{1}{\bar{v}_n} > 0$, this is equivalent to

$$-\frac{1}{\bar{v}_n}\tilde{d}_n^\top \bar{\mu}_n + \frac{1}{\bar{v}_n}\bar{\pi}_n^\top x_{a(n)} + \frac{1}{\bar{v}_n}\bar{\pi}_{n0}\theta_n \leq 0.$$

However, this is equivalent to

$$\tilde{d}_n^\top \mu_n^* - (\pi_n^*)^\top x_{a(n)} - \pi_{n0}^*\theta_n \leq 0$$

and by definition also to

$$\mathscr{L}_n^{i+1}(\pi_n^*, \pi_{n0}^*) - (\pi_n^*)^\top x_{a(n)} - \pi_{n0}^*\theta_n \leq 0,$$

which exactly corresponds to the Lagrangian cut (13). The reverse direction follows in a similar way. □

B.6 Proof of Lemma 3.26

Proof. Suppose that condition (19) is satisfied. Then there exist some $(\tilde{x}_{a(n)}, \tilde{\theta}_n) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1})) - (x_{a(n)}^i, \theta_n^i)$ and $\mu > 0$ such that $(u_n, u_{n0}) = \mu(\tilde{x}_{a(n)}, \tilde{\theta}_n)$. This implies $(\tilde{x}_{a(n)} + x_{a(n)}^i, \tilde{\theta}_n + \theta_n^i) \in \operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n^{i+1}))$. Therefore, the system

$$\left\{(\lambda_n, z_n) \;:\; \tilde{\theta}_n + \theta_n^i \geq c_n^\top \lambda_n, (\lambda_n, z_n) \in \operatorname{conv}(\mathcal{W}_n^{i+1}), z_n = \tilde{x}_{a(n)} + x_{a(n)}^i\right\}$$

is non-empty. However, this set is equivalent to

$$\Big\{(\lambda_n, z_n) \;:\; \frac{1}{\mu}u_{n0} \geq c_n^\top \lambda_n - \theta_n^i, (\lambda_n, z_n) \in \operatorname{conv}(\mathcal{W}_n^{i+1}), z_n - x_{a(n)}^i = \frac{1}{\mu}u_n\Big\}.$$

With choosing $\eta_n = \frac{1}{\mu} > 0$ it immediately follows that problem (16) is feasible. By $\eta_n \geq 0$, its optimal value is also bounded from below, hence it is finite. By LP duality this implies that Assumption 3 is satisfied. □

B.7 Proof of Lemma 4.1

Proof. Recall that $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$ can be expressed by problem (12). Moreover, according to Remark 3.3, under Assumption 1, the set $\operatorname{conv}(\mathcal{W}_n^{i+1})$ is a convex polyhedron. This implies that problem (12) can be reformulated as an LP where the state $x_{a(n)}$ appears in the RHS only. The assertion then follows from LP duality and the finite number of different dual extreme points. For $\overline{\text{co}}(Q_n)(\cdot)$ a very similar reasoning can be used. □

B.8 Proof of Theorem 4.2

Proof. We start with considering leaf nodes $n \in N \setminus \tilde{N}$, where $\Psi_m^{i+1} = \emptyset$ is fixed by definition. According to Lemma 4.1, $\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot)$ is a piecewise linear convex function, so $\operatorname{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1}))$ is polyhedral and has finitely many facets. As Theorem 3.27 (ii) is satisfied, all generated cuts are facet-defining. This means that only finitely many different cuts can be generated, and thus only finitely many different realizations of Ψ_n^{i+1} exist. In particular, after finitely many steps, for any $(x_{a(n)}^i, \theta_n^i)$ computed in the forward pass of NBD, a cut will have been generated that supports $\operatorname{epi}(\overline{\text{co}}(\underline{Q}_n^{i+1})(\cdot))$ at this point. Additionally, as no child nodes exist, we have $\underline{Q}_n^{i+1}(\cdot) \equiv Q_n(\cdot)$, and as $X_{a(n)} = \{0,1\}^{d_{a(n)}}$, we have $\overline{\text{co}}(Q_n)(x_{a(n)}) = Q_n(x_{a(n)})$ for all $x_{a(n)} \in X_{a(n)}$. This implies that the cut supports $\operatorname{epi}(Q_n)$ at $(x_{a(n)}^i, \theta_n^i)$ (*tight* cut).

For each realization of Ψ_n^{i+1}, the same reasoning as above can now be applied to the ancestor nodes of the leaf nodes. The assertion then follows by induction over the whole scenario tree. This implies that NBD terminates with an optimal policy for (MS-MILP). □

B.9 Proof of Lemma 4.3

Proof. For any $\tilde{\nu} \in \mathbb{N}$, the trial point $(\bar{x}_{a(n)}, \theta_n^{i_{\tilde{\nu}}})$ either satisfies $(\bar{x}_{a(n)}, \theta_n^{i_{\tilde{\nu}}}) \in \operatorname{epi}(\overline{\text{co}}(\underline{Q}_n^{i_{\tilde{\nu}}+1}))$, and by that $\theta_n^{i_{\tilde{\nu}}} = \bar{\theta}_n$, or it satisfies $\theta_n^{i_{\tilde{\nu}}} < \bar{\theta}_n$. In the first case, $(\bar{x}_{a(n)}, \theta_n^{i_{\tilde{\nu}}}) = (\bar{x}_{a(n)}, \bar{\theta}_n) = (\widehat{x}_{a(n)}^{i_{\tilde{\nu}}}, \widehat{\theta}_n^{i_{\tilde{\nu}}})$ and the assertion is trivially satisfied for all $\nu \geq \tilde{\nu}$.

In the second case, the trial point is separated from $\operatorname{epi}(\overline{\text{co}}(\underline{Q}_n^{i_{\tilde{\nu}}+1}))$ by a newly constructed Lagrangian cut. Therefore, the sequence $(\theta_n^{i_\nu})_{\nu\in\mathbb{N}}$ is monotonically increasing and bounded, thus converging. More precisely, by the above separation argument it converges to $\bar{\theta}_n$. It immediately follows that the distance between the trial point and $(\bar{x}_{a(n)}, \bar{\theta}_n)$ converges to zero:

$$\lim_{\nu\to\infty} \|\bar{x}_{a(n)} - x_{a(n)}^{i_\nu}, \bar{\theta}_n - \theta_n^{i_\nu}\|_* = \|\bar{x}_{a(n)} - \bar{x}_{a(n)}, \bar{\theta}_n - \theta_n^{i_\nu}\|_* = 0. \tag{27}$$

Since $(\bar{x}_{a(n)}, \bar{\theta}_n)$ is always feasible for the projection problem (15), this also implies that any sequence $(\widehat{x}_{a(n)}^{i_\nu}, \widehat{\theta}_n^{i_\nu})_{\nu\in\mathbb{N}}$ of (possibly non-unique) solutions of problem (15)

satisfies

$$\lim_{\nu\to\infty} \|\widehat{x}^{i_\nu}_{a(n)} - \bar{x}_{a(n)}, \widehat{\theta}^{i_\nu}_n - \theta^{i_\nu}_n\|_* = 0. \tag{28}$$

By using the triangle inequality together with (27) and (28) we obtain

$$\lim_{\nu\to\infty} \|\widehat{x}^{i_\nu}_{a(n)} - \bar{x}_{a(n)}, \widehat{\theta}^{i_\nu}_n - \bar{\theta}_n\|_* = 0.$$

From the definition of norms, the assertion follows. □

B.10 Proof of Lemma 4.4

Proof. We prove the result by contradiction. For that reason, we assume that there exists an infinite subsequence, indexed by $\ell \in \mathbb{N}$, such that for all ℓ there exists some $\widehat{k}^{\nu_\ell} \in \widehat{K}^{\nu_\ell}$ with $\widehat{k}^{\nu_\ell} \notin \overline{K}$. For simplicity, we assume that this facet index $\widehat{k}^{\nu_\ell}$ is the same for all ℓ, *i.e.*, $\widehat{k}^{\nu_\ell} \equiv: \tilde{k}$, even though this is not guaranteed. If it is not true, however, we can apply the same arguments after another restriction to subsequences for each possible facet, of which only finitely many exist.

Let the facet $F_{\tilde{k}}$ be described by the equation $q^\top x_{a(n)} + q_0\theta_n = r$, with appropriate coefficients $q_0, r \in \mathbb{R}$ and $q \in \mathbb{R}^{d_{a(n)}}$. Because of $\tilde{k} \in \widehat{K}^{\nu_\ell}$ for all ℓ, it follows that

$$q^\top \widehat{x}^{i_{\nu_\ell}}_{a(n)} + q_0\widehat{\theta}^{i_{\nu_\ell}}_n - r = 0$$

This immediately implies that

$$\lim_{\ell\to\infty} \left(q^\top \widehat{x}^{i_{\nu_\ell}}_{a(n)} + q_0\widehat{\theta}^{i_{\nu_\ell}}_n - r\right) = 0. \tag{29}$$

On the other hand, we know from Lemma 4.3 that $(\widehat{x}^{i_{\nu_\ell}}_{a(n)}, \widehat{\theta}^{i_{\nu_\ell}}_n)$ converges to $(\bar{x}_{a(n)}, \bar{\theta}_n)$, as for a convergent series every subsequence converges to the same limit. This implies

$$\begin{aligned}
&\lim_{\ell\to\infty} \left(q^\top \widehat{x}^{i_{\nu_\ell}}_{a(n)} + q_0\widehat{\theta}^{i_{\nu_\ell}}_n - r\right) \\
&= q^\top \lim_{\ell\to\infty} \left(\widehat{x}^{i_{\nu_\ell}}_{a(n)}\right) + q_0 \lim_{\ell\to\infty} \left(\theta^{i_{\nu_\ell}}_n\right) - r \\
&= q^\top \bar{x}_{a(n)} + q_0\bar{\theta}_n - r \\
&< 0.
\end{aligned} \tag{30}$$

The inequality follows from $\widehat{k}^{\nu_\ell} \notin \overline{K}$ for all $\ell \in \mathbb{N}$ and $(\bar{x}_{a(n)}, \bar{\theta}_n) \in \mathrm{epi}(\overline{\mathrm{co}}(\underline{Q}^{i+1}_n))$. Obviously, the results in (29) and (30) contradict each other, so our initial assumption must have been wrong. □

B.11 Proof of Lemma 4.6

Proof. Since at least x^2 satisfies $x^2 \in \mathrm{relint}(\mathfrak{F})$, there exists some $\lambda > 1$ such that the point $x^3 := x^1 + \lambda(x^2 - x^1) = (1-\lambda)x^1 + \lambda x^2$ is contained in $\mathfrak{F}$ as well. We can now prove the assertion by contradiction. Let α and β denote the intercept and the slope of the cut supporting S at x^2. Then, we have $\alpha + \beta^\top x^2 = 0$. Now assume that the cut does not support x^1, which implies $\alpha + \beta^\top x^1 < 0$, as $x^1 \in S$. We obtain

$$\alpha + \beta^\top x^3 = \lambda(\alpha + \beta^\top x^2) + (1-\lambda)(\alpha + \beta^\top x^1) = (1-\lambda)(\alpha + \beta^\top x^1) > 0. \tag{31}$$

The first equation applies the definition of x^3, the second one follows from $\alpha+\beta^\top x^2 = 0$. The third one follows from $\alpha+\beta^\top x^1 < 0$ and $\lambda > 1$. The result in (31) implies that either $x^3 \notin S$ or that the cut is not valid for all $x \in S$. Both cases lead to a contradiction. □

B.12 Proof of Lemma 4.7

Proof. We assume that ν is sufficiently large, that is, $\nu \geq \check{\nu}$ for some $\check{\nu} \in \mathbb{N}$ satisfying $\check{\nu} \geq \tilde{\nu}$ from Lemma 4.3, $\check{\nu} \geq \widehat{\nu}$ from Lemma 4.4 and $\check{\nu} \geq \bar{\nu}$ from Lemma 4.5.

If for all $\nu \geq \check{\nu}$ we have $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n) = (\bar{x}_{a(n)}, \bar{\theta}_n)$, then by Corollary 3.18 the assertion follows immediately. Therefore, we assume that this is not true, and distinguish different cases.

`Case 1.` Let $|\overline{K}| = 1$, *i.e.*, $(\bar{x}_{a(n)}, \bar{\theta}_n) \in \operatorname{int}(F_{\bar{k}})$ with $\overline{K} = \{\bar{k}\}$. From Lemma 4.4 it follows $\widehat{K}^\nu = \{\bar{k}\}$ for all $\nu \geq \check{\nu}$, thus $(\widehat{x}^{i}_{a(n)}, \widehat{\theta}^{i_\nu}_n) \in \operatorname{int}(F_{\bar{k}})$ as well. According to Corollary 3.18, for each ν, the obtained deep Lagrangian cut supports $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n))$ at $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$. The assertion then follows from Lemma 4.6.

`Case 2.` Let $|\overline{K}| > 1$, *i.e.*, $(\bar{x}_{a(n)}, \bar{\theta}_n) \in \operatorname{bd}(F_k)$ for all $k \in \overline{K}$. For any $\nu \geq \check{\nu}$ there are two possible sub-cases.

`Sub-case i).` The solution $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ to the projection problem (15) satisfies $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n) \in \operatorname{int}(F_k)$ for some $k \in \bar{K}$. Then, the assertion follows from Lemma 4.6.

`Sub-case ii).` The solution $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ to the projection problem (15) satisfies $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n) \in \operatorname{bd}(F_{k_\nu})$ for all $k^\nu \in \widehat{K}^\nu$, with $\widehat{K}^\nu \subseteq \overline{K}$. Then, $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ and $(\bar{x}_{a(n)}, \bar{\theta}_n)$ are still located on a joint sub-facet (face) $\mathfrak{F}$ of $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n))$. Additionally, Lemma 4.5 implies that for all $\nu \geq \check{\nu}$, $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n)$ is not a vertex of $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}_n))$. Therefore, we have $(\widehat{x}^{i_\nu}_{a(n)}, \widehat{\theta}^{i_\nu}_n) \in \operatorname{relint}(\mathfrak{F})$. Again, the assertion follows from Lemma 4.6. □

B.13 Proof of Theorem 4.8

Proof. We start with considering leaf nodes $n \in N \setminus \tilde{N}$, where $\Psi^{i+1}_m = \emptyset$ is fixed by definition. For any fixed $x_{a(n)} \in X_{a(n)}$, according to Lemma 4.7, after finitely many steps, a cut supporting $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}^{i+1}_n)(\cdot))$ at $(x^i_{a(n)}, \theta^i_n)$ is obtained. Since $\underline{Q}^{i+1}_n(\cdot) \equiv Q_n(\cdot)$ for the leaf nodes and $\overline{\operatorname{co}}(Q_n)(x_{a(n)}) = Q_n(x_{a(n)})$ for all $x_{a(n)} \in X_{a(n)} = \{0,1\}^{d_{a(n)}}$, this implies that the cut supports $\operatorname{epi}(Q_n)$ at $(x^i_{a(n)}, \theta^i_n)$ (*tight* cut). As $X_{a(n)}$ is a finite set and as Assumption 4 is satisfied, it follows (i) that only finitely many different cuts can be generated, and thus that only finitely many different realizations of Ψ^{i+1}_n exist, (ii) that after finitely many steps these realizations become exact at all x^i_n computed in the forward pass. Note that not all of these realizations have to be generated in NBD, though, but that it is also possible that an optimal x^*_n has been reached before (all possible) cuts have been generated at $x_n \neq x^*_n$.

For each realization of Ψ^{i+1}_n, the same reasoning as above can now be applied to the ancestor nodes of the leaf nodes. The assertion then follows by induction over the whole scenario tree. □

C Nested Benders Decomposition

In Algorithm 1, we provide a description of the NBD algorithm.

Algorithm 1 Nested Benders Decomposition with the new cut generation framework

Require: Scenario tree $\mathcal{T} = (\mathcal{N}, \mathcal{E})$, tolerance $\varepsilon > 0$, normalization functions $g_n(\cdot)$ for all $n \in \mathcal{N}$.

Initialization: bounds $\underline{v}^0 \leftarrow -\infty$, $\overline{v}^0 \leftarrow +\infty$, initial approximations Ψ_n^1 for all $n \in \mathcal{N}$ and iteration counter $i \leftarrow 0$.
while $\overline{v}^0 - \underline{v}^0 > \varepsilon$ **do**
    Set $i \leftarrow i + 1$.
    Solve subproblem (6) for the root node r to obtain a lower bound $\underline{v}^i$. Store the components $\left(x_r^i, y_r^i, (\theta_m^i)_{m \in \mathcal{C}(r)}\right)$ of the solution.
    for stages $t = 1, \ldots, T$ **do**
        for nodes $n \in \mathcal{N}_t$ **do**
            Solve subproblem (6) associated with function $\underline{Q}_n^i(x_{a(n)}^i)$ and store the components $\left(x_n^i, y_n^i, (\theta_m^i)_{m \in \mathcal{C}(n)}\right)$ of the solution. ▷ Forward pass
        end for
    end for
    Obtain an upper bound as $\overline{v}^i = \sum_{n \in \mathcal{N}} f_n(x_n^i, y_n^i)$.
    for stages $t = T, \ldots, 2$ **do** ▷ Backward pass
        for nodes $n \in \mathcal{N}_t$ **do**
            for children $m \in \mathcal{C}(n)$ **do**
                Solve the normalized Lagrangian dual (14) for (x_n^i, θ_m^i) and $g_n(\cdot)$ to compute a cut according to formula (13).
                Update Ψ_m^i to Ψ_m^{i+1} in node n using this cut.
            end for
        end for
    end for
end while

D Computational Results

D.1 CLSP with Binarization

The full computational results for our experiments of CLSP with state binarization are depicted in Tables 3-6. The table columns contain the number of stages, the used cut generation approach, the best lower bound obtained by SDDiP, a simulated statistical upper bound computed after termination of SDDiP (we report the upper limit of the computed confidence interval), the number of iterations, the time in seconds, the average time per iteration and the average number of iterations required in the level bundle method to solve the Lagrangian dual per iteration. We should note that in some cases, the simulation did not yield an upper bound estimate due to numerical issues.

In Table 3 we present the results for instances with $T = 4$, $T = 6$ and $T = 10$ stages and a time limit of 3 hours, 4 hours and 5 hours respectively. For $T = 4$, we also solve the deterministic equivalent with Gurobi for comparison.

In Table 4 we present the results for $T = 16$ stages and 8 hours of run time.

In Table 5 we present the results for runs that combine `SB` and different types of Lagrangian cuts.

In Table 6 we present the results for runs that use the Chen-Luedtke approach for a dual space restriction, again combining `SB` and different types of Lagrangian cuts.

D.2 CLSP without Binarization

For our tests of CLSP without state binarization, the full results are stated in Table 7.

Table 3: SDDiP results for CLSP with state binarization for $T = 4$, $T = 6$ and $T = 10$.

Method	Best LB	Stat. UB	Gap [%]	# Iter.	Time [s]	Time/It [s]	Lag-It/It
			$\boldsymbol{T = 4}$				
Det. Equiv.	1503.0	1542.3	3				
`B` (single)	803.2	1595.5	50	85	6	0	-
`B` (multi)	804.5	1602.6	50	59	5	0	-
`SB` (single)	1155.8	**1572.9**	27	32	16	1	-
`SB` (multi)	1187.1	1608.7	26	73	49	1	-
`L` (single)	681.8	1766.4	61	109	10800	99	49
`L` (multi)	702.6	1736.6	60	27	10908	404	51
ℓ^1-deep	1478.3	1582.5	**7**	39	11196	272	67
$\ell^{1\infty}$-deep	1462.9	1649.4	11	38	10944	288	63
ℓ^{∞}-deep	660.5	1854.1	64	43	11160	259	30
ℓ^{∞}-deep (MNC)	1049.0	1720.7	39	54	10944	203	77
`LN-Mid`	**1502.2**	-	-	38	11088	351	80
`LN-In-Out`	1017.0	1742.3	42	21	12024	573	97
`LN-Eps`	1499.0	-	-	28	11556	413	116
`LN-Relint`	1500.4	1621.2	8	33	11088	336	88
			$\boldsymbol{T = 6}$				
`B` (single)	1354.5	2854.2	53	479	69	0	-
`B` (multi)	1355.0	2889.4	53	165	37	0	-
`SB` (single)	2077.9	**2825.1**	26	46	41	1	-
`SB` (multi)	2093.4	2838.7	26	251	517	2	-
`L` (single)	669.7	3095.0	78	90	14580	162	50
`L` (multi)	682.5	3111.7	78	24	14688	612	130
ℓ^1-deep	2414.1	2975.4	19	34	14508	427	62
$\ell^{1\infty}$-deep	2375.8	2984.5	20	35	15012	429	59
ℓ^{∞}-deep (MNC)	1411.1	3003.5	53	48	14868	310	58
`LN-Mid`	**2500.3**	2892.9	**14**	24	14904	621	101
`LN-In-Out`	1162.2	3056.4	62	18	14940	830	97
`LN-Eps`	2454.2	2865.1	14	20	14580	729	135
`LN-Relint`	2477.0	2871.8	**14**	22	14616	664	108
			$\boldsymbol{T = 10}$				
`B` (single)	2165.3	5120.8	58	136	26	0	-
`B` (multi)	2183.6	5168.8	58	334	197	0	-
`SB` (single)	3377.1	4949.8	32	89	122	1	-
`SB` (multi)	3472.9	**4942.9**	30	206	1066	5	-
`L` (single)	616.4	5466.6	89	69	18144	263	51
`L` (multi)	682.4	5443.1	88	20	18720	936	54
ℓ^1-deep	3782.5	-	-	38	19260	507	47
$\ell^{1\infty}$-deep	3862.7	-	-	37	18504	500	47
ℓ^{∞}-deep (MNC)	2004.3	5256.8	62	44	18576	422	43
`LN-Mid`	**4222.2**	5168.9	**18**	18	18216	1012	110
`LN-Eps`	3642.8	5223.9	30	18	18360	1020	144
`LN-Relint`	4145.1	-	-	17	20232	1190	119

Table 4: SDDiP results for CLSP with state binarization for $T = 16$.

Method	Best LB	Stat. UB	Gap [%]	# Iter.	Time [s]	Time/Iter [s]	Lag-Iter/Iter
			$\boldsymbol{T = 16}$				
`B` (single)	3917.5	9012.6	57	224	74	0	-
`B` (multi)	3937.2	9008.2	56	254	274	1	-
`SB` (single)	5913.7	**8673.6**	32	194	493	3	-
`SB` (multi)	6030.2	8676.8	31	585	9252	16	-
`L` (single)	607.4	9524.5	94	63	28836	458	51
`L` (multi)	651.6	9444.7	93	19	31176	1641	55
ℓ^1-deep	3547.5	9756.3	64	42	28944	689	34
$\ell^{1\infty}$-deep	3353.0	9997.3	67	37	31248	845	16
ℓ^{∞}-deep (MNC)	2563.3	9917.7	74	52	29664	571	39
`LN-Mid`	**6910.7**	9125.9	**24**	18	32112	1784	119
`LN-Eps`	6014.5	9195.2	35	16	32040	2003	143
`LN-Relint`	6105.6	9085.9	33	15	29772	1985	123

Table 5: SDDiP results for CLSP using Lagrangian cuts combined with SB cuts.

Method	Best LB	Stat. UB	Gap [%]	# Iter.	Time [s]	Time/Iter [s]	Lag-Iter/Iter
			$\boldsymbol{T = 10}$				
`L` (single)	3475.5	**5003.9**	31	132	18072	136	50
`L` (multi)	3483.0	5014.4	31	51	18360	360	51
ℓ^1-deep	3804.1	5161.6	26	35	19656	562	88
$\ell^{1\infty}$-deep	3914.2	5233.7	25	36	19656	546	83
ℓ^{∞}-deep (MNC)	3551.5	5042.8	30	49	18108	340	180
`LN-Mid`	4052.9	5130.9	21	31	20484	661	115
`LN-Eps`	**4087.7**	5092.9	**20**	31	21096	681	141
`LN-Relint`	4034.3	5135.4	21	31	19944	643	119
			$\boldsymbol{T = 16}$				
`L` (single)	5907.1	**8728.4**	32	136	28944	213	47
`L` (multi)	5922.0	8750.8	32	52	30096	579	48
ℓ^1-deep	6477.5	9078.4	28	33	29916	907	90
$\ell^{1\infty}$-deep	6417.4	9024.6	29	34	30708	903	87
ℓ^{∞}-deep (MNC)	5991.3	8889.8	33	45	28800	640	191
`LN-Mid`	6675.7	9102.9	27	29	31680	1092	127
`LN-Eps`	6705.0	8947.8	**25**	29	30528	1053	146
`LN-Relint`	**6705.3**	8995.5	26	29	31284	1079	134

Table 6: SDDiP results for CLSP with $T = 10$ and $T = 16$ using the `CL` approach.

Method	Best LB	Stat. UB	Gap [%]	# Iter.	Time [s]	Time/Iter [s]	Lag-Iter/Iter
			$\boldsymbol{T = 10}$				
γ	4395.5	5224.4	**16**	74	18396	249	17
ℓ^1-deep	4344.5	-	-	53	18108	341	41
ℓ^∞-deep (MNC)	4249.9	5270.7	19	34	18216	536	383
`LN-Mid`	**4400.0**	**5217.6**	**16**	52	18288	352	34
`LN-Eps`	4389.3	5240.2	**16**	52	18684	359	38
`LN-Relint`	4371.4	5235.3	17	51	18036	354	34
			$\boldsymbol{T = 16}$				
γ	**7565.5**	9067.7	**17**	65	29160	449	18
ℓ^1-deep	7254.2	9095.4	20	49	29160	589	42
ℓ^∞-deep (MNC)	7248.9	9147.6	21	33	29808	903	432
`LN-Mid`	7521.0	9061.8	**17**	52	29340	611	38
`LN-Eps`	7510.8	9080.5	**17**	52	29232	622	40
`LN-Relint`	7543.3	**9051.4**	**17**	49	30204	617	38

Table 7: SDDiP results for CLSP with 10 state variables and no binary approximation.

Method	Best LB	Stat. UB	Gap [%]	# Iter.	Time [s]	Time/Iter [s]	Lag-Iter/Iter
$T = 16$, One type of cut							
`B` (single)	15190	29701	49	193	57	0	-
`B` (multi)	15199	29747	49	207	189	1	-
`SB` (single)	20373	29601	31	111	211	2	-
`SB` (multi)	21045	**29343**	28	281	3123	11	-
`L` (single)	13676	32003	57	84	10832	129	42
`L` (multi)	19199	31437	39	27	10971	406	45
ℓ^1-deep	23486	30367	23	34	11061	325	30
$\ell^{1\infty}$-deep	**23887**	30111	**21**	32	11472	359	33
ℓ^∞-deep (MNC)	23224	30074	23	26	10839	417	48
`LN-Mid`	19820	30450	35	19	11922	628	88
`LN-Eps`	22027	30599	28	22	10820	492	63
`LN-Relint`	8857	31188	72	19	11484	604	96
`LN-Conv`(50)	22588	30438	26	16	11132	696	98
`LN-Conv`(75)	22931	30206	24	17	10829	637	88
`LN-Conv`(90)	23215	30235	23	19	11342	597	79
`LN-Conv`(99)	22639	30374	26	22	11025	501	62
$T = 16$, Combination with `SB` cuts							
`L` (single)	21129	**30040**	30	92	10809	118	42
`L` (multi)	22271	30540	27	37	11312	306	44
ℓ^1-deep	**23235**	30342	**23**	35	11571	331	51
$\ell^{1\infty}$-deep	23197	30049	**23**	35	11364	325	49
ℓ^∞-deep (MNC)	22975	30210	24	36	11025	306	92
`LN-Mid`	22638	30413	26	30	12231	408	107
`LN-Eps`	22798	30382	25	32	10800	338	66
`LN-Relint`	22814	30303	25	29	11068	382	113
`LN-Conv`(50)	22641	30255	25	30	11941	398	110
`LN-Conv`(75)	22804	30120	24	31	12145	392	96
`LN-Conv`(90)	23162	30097	**23**	32	11590	362	79
`LN-Conv`(99)	23152	30118	**23**	33	11833	359	67
$T = 16$, Combination with `SB` cuts, `CL` approach							
γ	23715	**29894**	**21**	72	11028	153	23
ℓ^1-deep	**24036**	30231	**21**	48	11006	229	51
ℓ^∞-deep (MNC)	23803	30253	**21**	49	11051	226	82
`LN-Mid`	23114	30184	23	36	11118	309	98
`LN-Eps`	23566	30252	22	41	10822	264	65
`LN-Relint`	22938	30096	24	36	11288	314	99
`LN-Conv`(50)	22817	30344	25	31	11870	383	91
`LN-Conv`(75)	22701	30287	25	31	10974	354	82
`LN-Conv`(90)	22951	30344	24	32	10838	339	71
`LN-Conv`(99)	22902	30085	24	33	11084	336	61

References

[1] S. Ahmed, F. G. Cabral, and B. Freitas Paulo da Costa. Stochastic Lipschitz dynamic programming. *Mathematical Programming*, 191:755–793, 2022.

[2] D. Ávila, A. Papavasiliou, and N. Löhndorf. Batch learning SDDP for long-term hydrothermal planning. *IEEE Transactions on Power Systems*, 39(1):614–627, 2024.

[3] E. Balas and P. L. Ivanescu. On the generalized transportation problem. *Management Science*, 11(1):188–202, 1964.

[4] D. P. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[5] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: a fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[6] J. R. Birge. Solution methods for stochastic dynamic linear programs. Technical report, Stanford University CA Systems Optimization Lab, 1980.

[7] J. R. Birge and F. Louveaux. *Introduction to stochastic programming*. Springer Series in Operations Research and Financial Engineering. Springer Science & Business Media, 2nd edition, 2011.

[8] E. A. Boyd. Fenchel cutting planes for integer programs. *Operations Research*, 42(1):53–64, 1994.

[9] R. Brandenberg and P. Stursberg. Refined cut selection for benders decomposition: applied to network capacity expansion problems. *Mathematical Methods of Operations Research*, 94:383–412, 2021.

[10] F. Cadoux. Computing deep facet-defining disjunctive cuts for mixed-integer programming. *Mathematical Programming, Ser. A*, 122:197–223, 2010.

[11] R. Chen and J. Luedtke. On generating Lagrangian cuts for two-stage stochastic integer programs. *INFORMS Journal on Computing*, 2022.

[12] M. Conforti and L. A. Wolsey. "Facet" separation with one linear program. *Mathematical Programming, Ser. A*, 178:361–380, 2019.

[13] G. Cornuéjols and C. Lemaréchal. A convex-analysis perspective on disjunctive cuts. *Mathematical Programming, Ser. A*, 106:567–586, 2006.

[14] S. S. Dey and M. Molinaro. Theoretical challenges towards cutting-plane selection. *Mathematical Programming*, 170:237–266, 2018.

[15] O. Dowson and L. Kapelevich. SDDP.jl: a Julia package for stochastic dual dynamic programming. *INFORMS Journal on Computing*, 33(1):27–33, 2020.

[16] I. Dunning, J. Huchette, and M. Lubin. JuMP: a modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi: 10.1137/15m1020575.

[17] M. Fischetti, D. Salvagnin, and A. Zanette. A note on the selection of Benders' cuts. *Mathematical Programming, Ser. B*, 124:175–182, 2010.

[18] C. Füllner, X. A. Sun, and S. Rebennack. On Lipschitz regularization and Lagrangian cuts in multistage stochastic mixed-integer linear programming. Preprint, 2024.

[19] A. M. Geoffrion. Lagrangean relaxation for integer programming. In M.L. Balinski, editor, *Approaches to integer programming*, pages 82–114. Springer Berlin Heidelberg, Berlin, Heidelberg, 1974. ISBN 978-3-642-00740-8.

[20] M. Guignard. Lagrangean relaxation. *TOP*, 11(2):151–228, 2003.

[21] M. Hosseini and J. G. Turner. Deepest cuts for Benders decomposition. Preprint on arXiv, 2021.

[22] T.L. Magnanti and R.T. Wong. Accelerating benders decomposition: algorithmic enhancement and model selection criteria. *Operations Research*, 29(3):464–484, 1981.

[23] N. Papadakos. Practical enhancements to the Magnanti–Wong method. *Operations Research Letters*, 36:444–449, 2008.

[24] M. V. F. Pereira and L. M. V. G. Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52(1-3):359–375, 1991.

[25] R. Rahmaniani, S. Ahmed, T. G. Crainic, M. Gendreau, and W. Rei. The Benders dual decomposition method. *Operations Research*, 68(3):878–895, 2020.

[26] K. Seo, S. Joung, C. Lee, and S. Park. A closest Benders cut selection scheme for accelerating the Benders decomposition algorithm. *INFORMS Journal on Computing*, pages 1–24, 2022.

[27] H. D. Sherali and B. J. Lunday. On generating maximal nondominated Benders cuts. *Annals of Operations Research*, 210:57–72, 2013.

[28] P. M. Stursberg. *On the mathematics of energy system optimization*. PhD thesis, Technische Unviersität München, 2019.

[29] W. W. Trigeiro, L. J. Thomas, and J. O. McClain. Capacitated lot sizing with setup times. *Management Science*, 35(3):353–366, 1989.

[30] S. Zhang and X. A. Sun. Stochastic dual dynamic programming for multistage stochastic mixed-integer nonlinear optimization. *Mathematical Programming*, 2022.

[31] J. Zou, S. Ahmed, and X. A. Sun. Stochastic dual dynamic integer programming. *Mathematical Programming*, 175:461–502, 2019.