Normalization of ReLU Dual for Cut Generation in Stochastic Mixed-Integer Programs

Akul Bansal and Simge Küçükyavuz

Industrial Engineering and Management Sciences
Northwestern University
akul@u.northwestern.edu, simge@northwestern.edu

February 5, 2026

Abstract

We study the Rectified Linear Unit (ReLU) dual, an existing dual formulation for stochastic programs that reformulates non-anticipativity constraints using ReLU functions to generate tight, non-convex, and mixed-integer representable cuts. While this dual reformulation guarantees convergence with mixed-integer state variables, it admits multiple optimal solutions that can yield weak cuts. To address this issue, we propose normalizing the dual in the extended space to identify solutions that yield stronger cuts. We prove that the resulting normalized cuts are tight and Pareto-optimal in the original state space. We further compare normalization with existing regularization-based approaches for handling dual degeneracy and explain why normalization offers key advantages. In particular, we show that normalization can recover any cut obtained via regularization, whereas the converse does not hold. Computational experiments demonstrate that the proposed approach outperforms existing methods by consistently yielding stronger cuts and reducing solution times on harder instances.

1 Introduction

We consider a multistage stochastic integer program (MSIP) defined on a scenario tree $\mathcal{T}$ over stages $t = 1, \ldots, T$. Each node $n \in \mathcal{T}$ corresponds to a realization of uncertainty at a particular stage, and the set of nodes at stage t is denoted by N_t. For any node n, let $a(n)$ denote its unique ancestor, $C(n)$ its set of children, and q_{nm} the transition probability from node n to a child node $m \in C(n)$. At each node n, a d_n-dimensional state variable vector $x_n \in X_n \subseteq \mathbb{R}^{d_n}$ connects successive stages, and a local variable vector $y_n \in Y_n$ is specific to the subproblem at that node. The sets X_n and Y_n may include integrality restrictions. Given the state $x_{a(n)}$ inherited from its ancestor $a(n)$, the feasible decisions at node n are given by the polyhedron $H_n(x_{a(n)}) = \Big\{(x_n, y_n) \;\Big|\; T_n x_{a(n)} + W_n x_n + \overline{W}_n y_n = b_n\Big\}$, where $T_n, W_n, \overline{W}_n$ are technology and recourse matrices, and b_n is a right-hand side vector of appropriate dimension. The MSIP can then be formulated as:

$$\min_{x_1, y_1} \left\{ f_1(x_1, y_1) + \sum_{m \in C(1)} q_{1m}\, Q_m(x_1) \;:\; (x_1, y_1) \in H_1(x_0) \cap (X_1 \times Y_1) \right\}, \tag{1}$$

where for any node n, the associated value function is defined recursively as

$$Q_n(x_{a(n)}) = \min_{x_n, y_n} \; f_n(x_n, y_n) + \sum_{m \in C(n)} q_{nm}\, Q_m(x_n), \tag{2a}$$

$$(x_n, y_n) \in H_n(x_{a(n)}) \cap (X_n \times Y_n). \tag{2b}$$

The objective function $f_n(\cdot)$ is linear, and the initial state x_0 is given *apriori*. For final-stage nodes $n \in N_T$, the set of children nodes $C(n)$ is empty, and the expected future cost $\sum_{m \in C(n)} q_{nm}\, Q_m(x_n)$ is defined to be 0. The function $Q_n : \mathbb{R}^{d_{a(n)}} \to \mathbb{R} \cup \{+\infty\}$ takes value ∞ whenever the feasible set $H_n(x_{a(n)})$ is empty; we denote by $\mathrm{dom}(Q_n)$ the set of incumbent states $x_{a(n)}$ for which the subproblem Q_n is feasible.

Although MSIPs admit a deterministic equivalent mixed-integer programming (MIP) formulation, the number of variables and constraints grows exponentially with the size of the scenario tree. This quickly overwhelms off-the-shelf solvers and motivates the use of decomposition methods for large-scale instances (see Romeijndersa et al. 2025, Küçükyavuz and Sen 2017 for comprehensive reviews). A standard approach to decomposing the value function $Q_n(\cdot)$, is to reformulate the subproblem (2) as

$$Q_n\left(x_{a(n)}\right) = \min_{\substack{x_n, y_n, z_n, \\ (\theta_m)_{m \in C(n)}}} \; f_n(x_n, y_n) + \sum_{m \in C(n)} q_{nm}\, \theta_m, \tag{3a}$$

$$(x_n, y_n) \in H_n(z_n) \cap (X_n \times Y_n), \tag{3b}$$

$$(x_n, \theta_m) \in \mathrm{epi}(Q_m), \qquad \forall m \in C(n), \tag{3c}$$

$$z_n = x_{a(n)}, \tag{3d}$$

$$z_n \in Z_{a(n)} \supseteq X_{a(n)}, \tag{3e}$$

where the set $\mathrm{epi}(Q_m)$ is defined as follows:

$$\mathrm{epi}(Q_m) = \{(x_{a(m)}, \theta_m) \in \mathbb{R}^{d_{a(m)}} \times \mathbb{R} : \theta_m \geq Q_m(x_{a(m)}), x_{a(m)} \in \mathrm{dom}(Q_m)\}. \tag{4}$$

When needed, we write $\mathrm{epi}_S(p)$ to denote the epigraph of function p restricted to the set S. For instance, $\mathrm{epi}_{\mathrm{dom}(Q_m)}(Q_m)$ (or equivalently $\mathrm{epi}(Q_m)$ on $\mathrm{dom}(Q_m)$) denotes the epigraph in (4).

Because the value function $Q_m(\cdot)$ is not available in closed form, we approximate its epigraph $\mathrm{epi}(Q_m)$ with a polyhedral set Ψ_m which is iteratively refined by adding cuts of the form $h^i_m(x_n, \theta_m) \geq 0$. At a given iteration i, the inequality $h^i_m(\cdot, \cdot) \geq 0$ serves either as an optimality cut, yielding a valid lower bound on $Q_m(\cdot)$, or as a feasibility cut that approximates the domain $\mathrm{dom}(Q_m)$. In the nested Benders' method (Birge 1985), these cuts are computed by traversing the entire scenario tree, whereas sampling-based variants construct them using subsets of scenario paths at each iteration (Pereira and Pinto 1991, Zou et al. 2019, Füllner and Rebennack 2025).

To facilitate cut generation, it is common to introduce local variables z_n together with copy constraints $z_n = x_{a(n)}$, and relax the domain with $z_n \in Z_{a(n)} \supseteq X_{a(n)}$. For instance, Benders cuts (Benders 1962) can be derived from the optimal value of the linear programming (LP) relaxation of subproblem (3) and the corresponding LP dual solution associated with the copy constraints (3d). When the value function is convex polyhedral, it admits an exact representation using finitely many Benders cuts (see Rahmaniani et al. 2017 for a review on classical Benders decomposition). In contrast, when decision variables are integer or binary, the value function $Q_n(\cdot)$ is typically nonlinear and nonconvex (Blair and Jeroslow 1982). For problems with purely integer or binary variables, several specialized decomposition techniques have been proposed, including disjunctive programming (Sen and Higle 2005, Sen and Sherali 2006, Qi and Sen 2017), Fenchel cuts (Ntaimo 2013), and

parametric Gomory cuts (Gade et al. 2014, Zhang and Küçükyavuz 2014). Additional structure has also been exploited in other settings, such as when the value function is monotone (Philpott et al. 2020) or Lipschitz continuous (Ahmed et al. 2022, Füllner et al. 2024a).

More recently, there has been growing interest in decomposition methods with convergence guarantees for general MSIPs with *mixed-integer* state variables. For example, van der Laan and Romeijnders 2024 study scaled cuts and show that the associated scaled-cut closures converge uniformly to the convex envelope of the expected recourse function. However, these cuts cannot be computed using scenario decomposition methods, and the computational performance of the multistage extension (Romeijnders and van der Laan 2024) has not been tested.

In contrast, Zou et al. 2019 propose an alternative approach based on Lagrangian cuts derived by solving the Lagrangian dual obtained from relaxing the copy constraints (3d). Unlike Benders cuts, which only recover the LP relaxation of the value function, Lagrangian cuts can recover the closed convex envelope $\overline{\text{co}}(Q_n)$ of Q_n (see Chen and Luedtke 2022 and Füllner et al. 2024a for more details). Moreover, Zou et al. 2019 show that when the state variables are binary, Lagrangian cuts are sufficient to ensure tightness for Q_n, a property that, in turn, yields finite convergence of the decomposition method. To ensure convergence with non-binary state variables, Zou et al. 2019 propose encoding them with auxiliary binary variables. For continuous state variables, however, this discretization can dramatically expand the state space, increasing the cost of solving the Lagrangian dual and potentially making the problem computationally prohibitive (see Zou et al. 2019, Füllner et al. 2024b, Yang and Yang 2025 for implementation details and discussion).

Deng and Xie 2024 strike a middle-ground approach by ensuring convergence for general MSIPs with mixed-integer state variables without inflating the state dimension. They propose the following reformulation of subproblem (3)

$$
\begin{aligned}
Q_n\left(x_{a(n)}\right) = \min_{\substack{x_n, y_n, z_n, \\ (\theta_m)_{m\in C(n)}}} \quad & f_n(x_n, y_n) + \sum_{m\in C(n)} q_{nm}\,\theta_m, \\
\text{s.t. } & (3b), (3c), (3e), && \text{(5a)} \\
& (z_{nk} - x_{a(n)k})^+ = 0, \quad (z_{nk} - x_{a(n)k})^- = 0, \qquad \forall k \in [d_{a(n)}]. && \text{(5b)}
\end{aligned}
$$

The notation $[d_{a(n)}]$ denotes the set $\{1, \dots d_{a(n)}\}$. The copy constraints in (3d) are written using ReLU functions in (5b), where $(x)^+ := \max\{x, 0\}$ and $(x)^- := \max\{-x, 0\}$. These expressions are linearized using auxiliary variables that remain local to each subproblem, thereby preserving the state dimension. Relaxing the resulting copy constraints (5b) yields the ReLU dual, where the number of dual variables is bounded by twice the state dimension. Notably, the authors establish strong duality for this formulation, and the resulting ReLU-dual solutions can therefore be used to generate tight cuts even with mixed-integer state variables, which ensures asymptotic convergence for general MSIPs.

A related approach by Yang and Yang 2025 also establishes asymptotic convergence for general MSIPs. Instead of using ReLU-based copy constraints in (5b), they use the original copy constraints $z_n = x_{a(n)}$ in a lifted space. The ReLU-based copy constraints in (5b) can be interpreted as inducing a partition of the state space: along each dimension, the space is divided into regions to the left and right of the incumbent point, and the overall partition is obtained via the Cartesian product across all dimensions. The lifting procedure of Yang and Yang 2025 generalizes this partition created by iteratively introducing binary state variables and expanding the state dimension.

All dual-based cut-generation schemes—whether based on the LP dual (as in classical Benders decomposition), the Lagrangian dual (Zou et al. 2019), the ReLU dual of Deng and Xie 2024, or the lifted Lagrangian dual of Yang and Yang 2025—share a common challenge. They admit multiple

optimal solutions, some of which can generate weak cuts. For instance, Bansal and Küçükyavuz 2024 show that coefficients of integer L-shaped cuts–known for their weak global approximation and resulting slow convergence—are one of the optimal solutions to the Lagrangian dual.

In the LP-dual/Benders setting, this issue has been studied extensively, and a range of alternative cut-generation rules has been proposed to mitigate weak cuts and improve separation. Magnanti and Wong 1981 introduce a two-step procedure to obtain Pareto-optimal Benders cuts. Fischetti et al. 2010 introduce a unified cut-generation framework, which yields both feasibility and optimality cuts through a single separation problem, and admits methodological flexibility in identifying unbounded rays for cut generation through the choice of normalization. Different normalization choices can yield cuts that are deep (Hosseini and Turner 2025) and facet-defining or Pareto-optimal (Brandenberg and Stursberg 2021). Füllner et al. 2024b extend the work of Fischetti et al. 2010 to the Lagrangian dual of the mixed-integer subproblems to obtain Lagrangian cuts with desirable properties.

For more general duals—such as the ReLU dual and the lifted Lagrangian dual—Deng and Xie 2024, Yang and Yang 2025 propose mitigation strategies that extend the ideas of Magnanti and Wong 1981 to their formulations. These strategies can be viewed as dual regularization: they seek to characterize the set of all optimal dual solutions and then select those that yield stronger cuts by optimizing a regularized objective over this set. The modified objective includes additional terms involving the dual variables, weighted by appropriately chosen objective coefficients. In particular, Deng and Xie 2024 approximate the set of optimal dual solutions using a linear program. They derive objective coefficients so that the regularized dual remains bounded. The cut-generating LP is computationally efficient but often results in weak cuts. In contrast, Yang and Yang 2025 characterize the set of optimal dual solutions exactly using a convex program. They propose objective coefficients that lead to cuts that are both tight and Pareto-optimal in the lifted space.

In this work, we propose an alternative approach to mitigating weak cuts resulting from dual degeneracy in methods designed for mixed-integer state variables. Our approach is based on normalization of the ReLU dual. This approach was introduced by Füllner et al. 2024b for the Lagrangian dual obtained by relaxing copy constraints $z_n = x_{a(n)}$ in MIP subproblems. In the normalized dual, additional constraints are imposed on the dual variables, which are shown to address the weakness of the original Lagrangian cuts arising from dual degeneracy. However, the original Lagrangian cuts, based on relaxing original copy constraints $z_n = x_{a(n)}$, only guarantee convergence when the state variables are binary (Zou et al. 2019). We extend the concept of normalization to the ReLU-based dual setting to ensure convergence with mixed-integer state variables and to obtain strong cuts.

We focus on the ReLU dual rather than the lifted dual of Yang and Yang 2025 because the latter introduces binary variables directly into the state space. This can lead to a significant expansion of the state space, substantially increasing the computational effort required to solve the lifted Lagrangian dual and generate the corresponding cuts. In contrast, the ReLU dual introduces auxiliary binary variables only as local variables to the subproblem, keeping the number of dual variables bounded by twice the state dimension in each iteration. This makes solving the dual significantly more efficient than the lifted approach. In our computational study, we observe that the time to solve the dual and the average number of dual iterations both increase with the state dimension. We also study the relationship between normalization and regularization and explain why normalization offers important advantages. In particular, we show that normalization can yield tight, Pareto-optimal cuts. To this end, we introduce a notion of Pareto-optimality defined in the original state space for non-linear ReLU cuts and prove that normalization yields Pareto-optimal cuts under this definition. Importantly, normalization provides additional flexibility: cuts need not be both tight and Pareto-optimal simultaneously. By appropriately choosing the normalization coefficients, one may enforce both properties, but this is not required. In contrast, the enhancement strategy proposed by Yang and Yang 2025 always produces cuts that are both tight and Pareto-optimal. This

lack of flexibility can lead to weaker overall approximations of the value function, a behavior that we also illustrate in our computational experiments.

The main contributions of this paper are as follows:

1. We extend the normalization framework of Füllner et al. 2024b to the ReLU-based Lagrangian dual of Deng and Xie 2024. This extension ensures asymptotic convergence for multistage stochastic integer programs with mixed-integer state variables while producing strong cuts.
2. We show that normalization resolves the issue of weak cuts resulting from degeneracy in the ReLU dual and identifies solutions that yield strong cuts. The strength of the cuts is established from two perspectives:
   (a) We introduce a notion of Pareto-optimality defined in the original state space for nonlinear ReLU cuts, and prove that normalization produces Pareto-optimal cuts under this definition.
   (b) We prove that there exists a choice of normalization coefficients that yields tight cuts at the current incumbent. Our existence result also provides insights into selecting these coefficients to obtain tight cuts.
3. We analyze the relationship between normalization and regularization for obtaining strong cuts. We show that any cut obtained by regularization can also be obtained by normalization, though the converse is not true. While we establish this relationship in the context of the ReLU dual, it has broader implications. For classical Benders cuts, previous work (Brandenberg and Stursberg 2021, Hosseini and Turner 2025) shows that both normalization and the regularization framework of Magnanti and Wong 1981 can attain Pareto-optimal cuts. However, when multiple such cuts exist, it is unclear whether normalization can attain the same cut as regularization. Our result resolves this question and proves a stronger claim: with an appropriate choice of normalization constraints, the coefficients of any Pareto-optimal cut obtained via a regularization method are also optimal in the normalization dual.
4. Finally, we provide a computational comparison of normalization and regularization approaches, showing that normalization consistently yields stronger cuts, while computational times improve for harder instances that cannot be effectively approximated by the weaker Benders cuts. We also make our code publicly available. To the best of our knowledge, it is the first open-source implementation of a cut-generation method with (asymptotic) convergence guarantees for general MSIPs. The code supports strong cut generation via both normalization and regularization, among other enhancements such as the alternating cut strategy of Angulo et al. 2016.

The remainder of this paper is organized as follows. In Section 2, we introduce normalization of the ReLU dual and prove that the resulting cuts are tight and Pareto-optimal. Section 3 revisits recently proposed regularization-based approaches and discusses their connection to normalization. Section 4 presents a computational comparison of normalization and regularization across two classes of problems. Finally, Section 5 summarizes our contributions and discusses directions for future work.

2 Normalization of the Dual Formulation

2.1 ReLU Dual and ReLU Cuts

We first review the ReLU dual and the associated ReLU cuts introduced by Deng and Xie 2024. Given an incumbent solution $\hat{x}_{a(n)}$, cuts are obtained by relaxing the copy constraints (5b), which leads to the following Lagrangian relaxation:

$$\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) := \min_{z_n \in Z_{a(n)}} \underline{Q}_n(z_n) + \sum_{k \in [d_{a(n)}]} \pi_{nk}^+ (z_{nk} - \hat{x}_{a(n),k})^+ + \sum_{k \in [d_{a(n)}]} \pi_{nk}^- (z_{nk} - \hat{x}_{a(n),k})^-, \tag{6}$$

where the approximate value function $\underline{Q}_n$ is obtained by replacing $\mathrm{epi}(Q_m)$ in (3c) and (5a) with an approximation Ψ_m iteratively obtained from the cut-generation process. The corresponding Lagrangian dual problem is given by:

$$\max_{\pi_n^+, \pi_n^- \in \mathbb{R}^{d_{a(n)}}} \mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}). \tag{7}$$

We refer to this problem as the ReLU dual and use the superscript R to denote expressions associated with this formulation.

For any choice of dual multipliers $\pi_n^+, \pi_n^- \in \mathbb{R}^{d_{a(n)}}$, a ReLU cut that is valid for the epigraph $\mathrm{epi}_{Z_{a(n)}}(Q_n)$ is given by

$$\theta_n \geq \mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) - \sum_{k \in [d_{a(n)}]} \pi_{nk}^+ (z_{nk} - \hat{x}_{a(n),k})^+ - \sum_{k \in [d_{a(n)}]} \pi_{nk}^- (z_{nk} - \hat{x}_{a(n),k})^-. \tag{8}$$

Deng and Xie 2024 show that strong duality holds for the ReLU dual (7) under standard assumptions. Strong duality guarantees that a cut generated at the current incumbent is tight. A key advantage of the ReLU-based copy constraints is that they enable the construction of tight cuts even when the state variables are mixed-integer. As a result, the approach ensures asymptotic convergence for general mixed-integer stochastic programs with mixed-integer state variables.

The authors further discuss a linear reformulation of the ReLU cut (8). This reformulation models the ReLU functions $(z_{nk} - x_{a(n),k})^+$ and $(z_{nk} - x_{a(n),k})^-$ implicitly through auxiliary continuous variables w_{nk}^+, w_{nk}^- and binary variables r_{nk}. In particular, given known bounds on the state variables $x_{nk} \in [0, B_k]$, the ReLU cut can be expressed as the following MIP formulation:

$$\theta_n \geq \mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) - \sum_{k \in [d_{a(n)}]} \pi_{nk}^+ w_{nk}^+ - \sum_{k \in [d_{a(n)}]} \pi_{nk}^- w_{nk}^- \tag{9a}$$

$$w_{nk}^+ - w_{nk}^- = z_{nk} - \hat{x}_{a(n),k}, \qquad \forall k \in [d_{a(n)}], \tag{9b}$$

$$0 \leq w_{nk}^+ \leq (B_k - \hat{x}_{a(n),k}) r_{nk}, \qquad \forall k \in [d_{a(n)}], \tag{9c}$$

$$0 \leq w_{nk}^- \leq \hat{x}_{a(n),k} (1 - r_{nk}), \qquad \forall k \in [d_{a(n)}], \tag{9d}$$

$$r_n \in \{0,1\}^{d_{a(n)}}. \tag{9e}$$

This reformulation reveals an important connection between ReLU cuts and the original Lagrangian cuts in a lifted space, which we discuss in the next section.

For the rest of the paper, we make the following basic assumptions. First, we assume that sets X_n and Y_n are compact, and all problem data are rational. This ensures that feasible region

$H_n(x_{a(n)}) \cap (X_n \times Y_n)$ is compact for all $x_{a(n)} \in \mathbb{R}^{d_{a(n)}}$. Moreover, under this assumption, the set X_n can be shifted such that all state variables satisfy $x_{nk} \in [0, B_k]$. The data are assumed to be rational to ensure finite MIP-representations and because the convex hull of any MIP-representable set defined by rational data is a rational polyhedron (Meyer 1974). Second, we assume that the domain of the value function satisfies $\text{dom}(Q_n) = Z_{a(n)} = \prod_k [0, B_k]$. Together with (3e), this property guarantees relatively complete recourse, a standard assumption in stochastic programming. This requirement can always be enforced by adding continuous variables to polyhedron H_n and penalizing them in the objective. The set $Z_{a(n)}$ is typically taken as a superset of $X_{a(n)}$ to reduce the computational effort in solving the dual problem, but this choice affects the strength of the resulting cuts. For example, in Zou et al. 2019, binary restrictions on state variables are relaxed to the interval $[0, 1]$. We refer the reader to Füllner et al. 2024a for a detailed discussion of the various options for choosing $Z_{a(n)}$ and their implications.

2.2 Connection with the Original Lagrangian Cuts

To establish the connection between ReLU cuts and the original Lagrangian cuts, we lift the domain $Z_{a(n)}$ of the value function Q_n to a higher-dimensional space. Given an incumbent solution $\hat{x}_{a(n)}$, we define the lifted domain as:

$$Z^{lift}_{\hat{x}_{a(n)}} = \{(w_n^+, w_n^-) : \exists z_n \in Z_{a(n)} \text{ and } \exists r_n \in \{0,1\}^{d_{a(n)}} \text{ s.t. } (9b) - (9d)\}. \tag{10}$$

This lifted space corresponds to the auxiliary variables (w_n^+, w_n^-) used in the linear reformulation (9) of the ReLU cut. On this lifted space, we redefine the recourse function as:

$$\underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)}) := \underline{Q}_n(\hat{x}_{a(n)} + w_n^+ - w_n^-). \tag{11}$$

Note that $\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)})$, which allows us to reformulate the subproblem $\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$ with the standard copy constraints:

$$\begin{aligned} \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)}) = \min_{w_n^+, w_n^-} \quad & \underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)}) \\ \text{s.t.} \quad & (w_n^+, w_n^-) = (\mathbf{0}, \mathbf{0}), \\ & (w_n^+, w_n^-) \in Z^{lift}_{\hat{x}_{a(n)}}. \end{aligned} \tag{12}$$

By relaxing the copy constraints (12), we obtain the following Lagrangian relaxation:

$$\begin{aligned} \mathcal{L}_n^O(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) := \min_{w_n^+, w_n^-} \quad & \underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)}) + (\pi_n^+, \pi_n^-) \cdot \begin{pmatrix} w_n^+ \\ w_n^- \end{pmatrix} \\ \text{s.t.} \quad & (w_n^+, w_n^-) \in Z^{lift}_{\hat{x}_{a(n)}}. \end{aligned} \tag{13}$$

The superscript O in $\mathcal{L}_n^O$ denotes the original Lagrangian relaxation obtained by relaxing the copy constraints (12). The corresponding Lagrangian dual problem is:

$$\max_{\pi_n^+, \pi_n^-} \mathcal{L}_n^O(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) - (\pi_n^+, \pi_n^-) \cdot \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \tag{14}$$

For any dual multipliers (π_n^+, π_n^-), the resulting Lagrangian cut takes the form:

$$\theta_n \geq \mathcal{L}_n^O(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) - (\pi_n^+, \pi_n^-) \cdot \begin{pmatrix} w_n^+ \\ w_n^- \end{pmatrix}. \tag{15}$$

This transformation establishes the following key relationship:

Proposition 1 (Deng and Xie 2024). *The ReLU Lagrangian cut (8), generated at $\hat{x}_{a(n)}$ for $\text{epi}_{Z_{a(n)}}\left(\underline{Q}_n\right)$, corresponds to Lagrangian cut (15) generated at $(\mathbf{0},\mathbf{0})$ for the lifted set $\text{epi}_{Z^{lift}_{\hat{x}_{a(n)}}}\left(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})\right)$. Moreover, with optimal dual multipliers obtained by solving the ReLU dual (7), the Lagrangian cut (15) is tight at $(\mathbf{0},\mathbf{0})$ with respect to the lifted value function $\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})$.*

Although this result is stated without proof in the original paper, we provide a proof in Appendix A for completeness. The main takeaway of Proposition 1 is that any optimal solution of the ReLU dual (7) is also optimal for the lifted Lagrangian dual (14) and vice-versa. This means that properties of the original Lagrangian cuts can be applied to the transformed problem $\underline{Q}'_n$ over the lifted domain $Z^{lift}_{\hat{x}_{a(n)}}$.

2.3 Normalized ReLU Dual and Normalized ReLU cuts

Proposition 1 enables us to apply the normalization procedure of Füllner et al. 2024b to the transformed subproblem $\underline{Q}'_n$ in (11), domain $Z^{lift}_{\hat{x}_{a(n)}}$, and incumbent $(\mathbf{0},\mathbf{0})$. More precisely, given incumbent solution $(\mathbf{0},\mathbf{0})$ to subproblem $\underline{Q}'_n$ and approximation $\hat{\theta}_n$ of the subproblem cost, we consider the following feasibility version of the subproblem $\underline{Q}'_n(\mathbf{0},\mathbf{0};\hat{x}_{a(n)})$:

$$\begin{aligned}
\underline{v}_n((\mathbf{0},\mathbf{0}),\hat{\theta}_n) := \min \quad & 0 \\
& \underline{Q}'_n(w_n^+,w_n^-;\hat{x}_{a(n)}) \le \hat{\theta}_n && \text{(16a)}\\
& (w_n^+,w_n^-) = (\mathbf{0},\mathbf{0}) && \text{(16b)}\\
& (w_n^+,w_n^-) \in Z^{lift}_{\hat{x}_{a(n)}}.
\end{aligned}$$

Next, a normalized dual is obtained by dualizing the constraints (16a) and (16b) with the corresponding dual variables $\pi_{n0} \in \mathbb{R}$ and $\pi_n^+, \pi_n^- \in \mathbb{R}^{d_{a(n)}}$, respectively, and adding a constraint normalizing the dual variables. In particular, consider the following dual problem:

$$\underline{v}_n^{ND}((\mathbf{0},\mathbf{0}),\hat{\theta}_n) := \max_{\pi_n^+,\pi_n^-,\pi_{n0}} \left(\mathcal{L}_n(\pi_n^+,\pi_n^-,\pi_{n0};\hat{x}_{a(n)}) - (\pi_n^+,\pi_n^-)\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix} - \pi_{n0}\hat{\theta}_n\right), \tag{17a}$$

$$g_n(\pi_n^+,\pi_n^-,\pi_{n0}) \le 1, \pi_{n0} \ge 0, \tag{17b}$$

where function $g_n(\pi_n^+,\pi_n^-,\pi_{n0})$ in the normalization constraint is of the form $u_n^+\pi_n^+ + u_n^-\pi_n^- + u_{n0}\pi_{n0}$, where u_n^+, u_n^-, u_{n0} are given normalization coefficients, and the Lagrangian relaxation $\mathcal{L}_n(\pi_n^+,\pi_n^-,\pi_{n0};\hat{x}_{a(n)})$ is given as follows:

$$\mathcal{L}_n(\pi_n^+,\pi_n^-,\pi_{n0};\hat{x}_{a(n)}) = \min_{w_n^+,w_n^-}\{(\pi_n^+,\pi_n^-)\begin{pmatrix}w_n^+\\ w_n^-\end{pmatrix} + \pi_{n0}\left(\underline{Q}'_n(w_n^+,w_n^-;\hat{x}_{a(n)})\right), (w_n^+,w_n^-) \in Z^{lift}_{\hat{x}_{a(n)}}\}.$$

In Section 2.4, we discuss how to choose the normalization coefficients u_n^+, u_n^-, u_{n0} to obtain cuts with desired properties.

Based on an optimal solution $(\hat{\pi}_n^+,\hat{\pi}_n^-,\hat{\pi}_{n0})$ of the normalized dual (17), the resulting normalized cut is

$$\hat{\pi}_{n0}\theta_n \ge \mathcal{L}_n(\hat{\pi}_n^+,\hat{\pi}_n^-,\hat{\pi}_{n0};\hat{x}_{a(n)}) - (\hat{\pi}_n^+,\hat{\pi}_n^-)\begin{pmatrix}w_n^+\\ w_n^-\end{pmatrix}. \tag{18}$$

The validity of cut (18) for $\text{epi}_{Z^{lift}_{\hat{x}_{a(n)}}}(\underline{Q}'_n)$ follows from Lemma 3.7 of Füllner et al. 2024b. It is easy to see that this validity implies that the associated ReLU cut

$$\hat{\pi}_{n0}\theta_n \geq \mathcal{L}_n(\hat{\pi}^+_n, \hat{\pi}^-_n, \hat{\pi}_{n0}; \hat{x}_{a(n)}) - \sum_{k\in[d_{a(n)}]} \pi^+_{nk}(z_{nk} - \hat{x}_{a(n),k})^+ - \sum_{k\in[d_{a(n)}]} \pi^-_{nk}(z_{nk} - \hat{x}_{a(n),k})^-, \quad (19)$$

is also valid for the set $\text{epi}_{Z_{a(n)}}(\underline{Q}_n)$. Now, consider the incumbent $(\hat{x}_{a(n)}, \hat{\theta}_n)$ in the original space such that $\hat{\theta}_n < \underline{Q}_n(\hat{x}_{a(n)}) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$. Then, we seek a cut of the form (18) to cut off this incumbent. Our next result establishes that such a cut always exists.

Proposition 2. *If $\hat{\theta}_n \geq \underline{Q}_n(\hat{x}_{a(n)}) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$, then $\underline{v}^{ND}_n((\mathbf{0}, \mathbf{0}), \hat{\theta}_n) = 0$. Otherwise, there exists a cut of the form (18), such that the incumbent $((\mathbf{0}, \mathbf{0}), \hat{\theta}_n)$ violates this cut. In the original space, the incumbent $(\hat{x}_{a(n)}, \hat{\theta}_n)$, violates the corresponding ReLU cut of the form (19).*

Proof. We prove the two statements separately. First, we suppose that $\hat{\theta}_n \geq \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$ and show that $\underline{v}^{ND}_n((\mathbf{0}, \mathbf{0}), \hat{\theta}_n) = 0$. Let $\overline{\text{co}}(f)$ denote the closed convex envelop of function f and $\overline{\text{co}}(f(y))$ denote the function $\overline{\text{co}}(f)$ evaluated at y.

Since the closed convex envelope satisfies $\overline{\text{co}}\Big(\underline{Q}'_n((\mathbf{0}, \mathbf{0}); \hat{x}_{a(n)})\Big) \leq \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$, we have $((\mathbf{0}, \mathbf{0}), \hat{\theta}_n) \in \text{epi}_{Z^{lift}_{\hat{x}_{a(n)}}}\Big(\overline{\text{co}}\Big(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)})\Big)\Big)$.

Now, note that the normalization function g_n of the form $u^+_n\pi^+_n + u^-_n\pi^-_n + u_{n0}\pi_{n0}$ is homogeneous. Therefore, by Lemma 3.14 of Füllner et al. 2024b, we obtain $\underline{v}^{ND}_n((\mathbf{0}, \mathbf{0}), \hat{\theta}_n) = 0$.

Next, suppose that $\hat{\theta}_n < \underline{Q}_n(\hat{x}_{a(n)}) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$. By Theorem 3.5 and Remark 3.13 of Füllner et al. 2024b, it suffices to show that $\hat{\theta}_n < \overline{\text{co}}\Big(\underline{Q}'_n((\mathbf{0}, \mathbf{0}); \hat{x}_{a(n)})\Big)$, because then there exists a normalized Lagrangian cut of the form (18) that separates the incumbent.

To this end, Theorem 3.13 of Füllner et al. 2024a implies that $\overline{\text{co}}\Big(\underline{Q}'_n((\mathbf{0}, \mathbf{0}); \hat{x}_{a(n)})\Big) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$ whenever $(\mathbf{0}, \mathbf{0})$ is an extreme point of $Z^{lift}_{\hat{x}_{a(n)}}$. We now verify that $(\mathbf{0}, \mathbf{0})$ is an extreme point. Let $\lambda \in (0, 1)$ and let $(v^+_1, v^-_1), (v^+_2, v^-_2) \in Z^{lift}_{\hat{x}_{a(n)}}$ satisfy $\lambda(v^+_1, v^-_1) + (1 - \lambda)(v^+_2, v^-_2) = (\mathbf{0}, \mathbf{0})$. With known bounds on state variable $x_{nk} \in [0, B_k], k \in [d_{a(n)}]$, the domain $Z^{lift}_{\hat{x}_{a(n)}} \subseteq \mathbb{R}^{d_{a(n)}}_+ \times \mathbb{R}^{d_{a(n)}}_+$, so all components of $v^+_1, v^-_1, v^+_2, v^-_2$ are nonnegative. The above convex-combination equality therefore forces $(v^+_1, v^-_1) = (v^+_2, v^-_2) = (\mathbf{0}, \mathbf{0})$ proving that $(\mathbf{0}, \mathbf{0})$ is an extreme point of $Z^{lift}_{\hat{x}_{a(n)}}$. Consequently, $\overline{\text{co}}\Big(\underline{Q}'_n((\mathbf{0}, \mathbf{0}); \hat{x}_{a(n)})\Big) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$ and since $\hat{\theta}_n$ is strictly smaller than this value, a separating normalized cut exists.

Using linear reformulation in (9b)-(9e), this translates to a ReLU cut of the form (19) that cuts-off the incumbent $(\hat{x}_{a(n)}, \hat{\theta}_n)$. □

In Füllner et al. 2024b, the existence of a separating normalized Lagrangian cut is established under the condition $(\hat{x}_{a(n)}, \hat{\theta}_n) \notin \text{epi}_{Z_{a(n)}}\Big(\overline{\text{co}}(\underline{Q}_n)\Big)$. Proposition 2 strengthens this by showing that, a normalized cut of the form (19) exists whenever $(\hat{x}_{a(n)}, \hat{\theta}_n) \notin \text{epi}_{Z_{a(n)}}(\underline{Q}_n)$. Consequently, normalized ReLU cuts preserve the separation property needed for asymptotic convergence.

2.4 Normalization Coefficients and their Impact on Cut Quality

In this section, we discuss how to choose the normalization coefficients u_n^+, u_n^-, and u_{n0} in the normalization constraint (17b). This choice impacts cut quality, which we evaluate through two properties: Pareto-optimality and tightness at the incumbent. We first study these properties in the extended space and then translate the results back to the original state space.

2.4.1 Pareto-optimal Cuts

The concept of Pareto-optimal cuts is introduced in Magnanti and Wong 1981 for affine cuts.

Definition 1 (Pareto-optimal affine cut)**.** *A cut of the form $\theta_n \geq \ell^1 - (\pi^1)^\top x_{a(n)}$ dominates the cut $\theta_n \geq \ell^2 - (\pi^2)^\top x_{a(n)}$ if $\ell^1 - (\pi^1)^\top x_{a(n)} \geq \ell^2 - (\pi^2)^\top x_{a(n)}$ for all $x_{a(n)} \in X_{a(n)}$, with strict inequality for at least one point in $X_{a(n)}$. A valid cut is Pareto-optimal for reference set $\operatorname{epi}_{X_{a(n)}}(\underline{Q}_n)$ if no other valid cut dominates it.*

Pareto-optimality is always defined relative to a reference set. Importantly, while establishing Pareto-optimality on a larger set such as $\operatorname{epi}_{\operatorname{conv}(X_{a(n)})}(\overline{\operatorname{co}}(\underline{Q}_n))$ may be easier, it does not guarantee Pareto-optimality on the original epigraph $\operatorname{epi}_{X_{a(n)}}(\underline{Q}_n)$. We refer the reader to Figure 3.4 and Figure 3.5 of Stursberg 2019 for an example depicting the importance of the reference set in the definition of the Pareto-optimal cuts.

We now apply this concept to the normalized (and linear) cut (18) in the extended space. Our goal is to identify the choice of normalization coefficients u_n^+, u_n^-, u_{n0} in constraint (17b) that yields Pareto-optimal cuts. The following result provides this characterization. Throughout, $\operatorname{relint}(\cdot)$ denotes the relative interior of a set.

Proposition 3. *For all $(u_n^+, u_n^-, u_{n0}) \in \operatorname{relint}\Big(\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))) - (\mathbf{0},\mathbf{0},\hat{\theta}_n)\Big)$, any optimal solution $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ to the normalized dual (17) with $\hat{\pi}_{n0} > 0$ defines a Pareto-optimal cut of form (18) for the reference set $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})))$ on $\operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}})$.*

The proof follows immediately from Theorem 3.25 of Füllner et al. 2024b. In the standard literature, normalization coefficients that yield Pareto-optimal cuts or satisfy the requirement in Proposition 3 are referred to as *core points*. While identifying core points remains a challenge and typically requires solving an additional MIP subproblem (Magnanti and Wong 1981, Füllner et al. 2024b, Yang and Yang 2025), we show that when the incumbent solution $\hat{x}_{a(n)}$ satisfies $0 < \hat{x}_{a(n),k} < B_k$ for all $k \in [d_{a(n)}]$, a core point can be constructed efficiently without additional subproblem solves, as shown below.

Proposition 4. *Suppose that the incumbent solution $\hat{x}_{a(n)}$ satisfies $0 < \hat{x}_{a(n),k} < B_k$ for all $k \in [d_{a(n)}]$. Then any point (u_n^+, u_n^-, u_{n0}) with*

$$(u_{nk}^+, u_{nk}^- = u_{nk}^+) \in \operatorname{relint}\Big(\operatorname{conv}\Big\{(0,0),\ \big(B_k - \hat{x}_{a(n),k},\, 0\big),\ \big(0,\, \hat{x}_{a(n),k}\big)\Big\}\Big) \tag{20}$$

for all $k \in [d_{a(n)}]$, and $u_{n0} \in (\underline{Q}_n(\hat{x}_{a(n)}) - \hat{\theta}_n, \infty)$, defines a core point, that is, $(u_n^+, u_n^-, u_{n0}) \in \operatorname{relint}\Big(\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))) - (\mathbf{0},\mathbf{0},\hat{\theta}_n)\Big)$.

Proof. Consider any coefficients (u_n^+, u_n^-, u_{n0}) satisfying condition (20) with $u_{n0} \in (\underline{Q}_n(\hat{x}_{a(n)}) - \hat{\theta}_n, \infty)$. We prove that $(u_n^+, u_n^-, u_{n0}) \in \operatorname{relint}\Big(\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))) - (\mathbf{0},\mathbf{0},\hat{\theta}_n)\Big)$.

Since $\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})$ is defined on $Z^{lift}_{\hat{x}_{a(n)}}$, $\overline{\text{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$ is defined on $\text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$. Thus, it suffices to show that $(u_n^+, u_n^-) \in \text{relint}(\text{conv}(Z^{lift}_{\hat{x}_{a(n)}}))$ and $u_{n0} > \overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) - \hat{\theta}_n$.

We first show that the chosen u_n^+, u_n^- satisfy the relative interior requirement. To this end, we analyze the lifted domain $Z^{lift}_{\hat{x}_{a(n)}}$ defined in (10). For notational simplicity, we denote $Z^L := Z^{lift}_{\hat{x}_{a(n)}}$ and $Z := Z_{a(n)}$ in this proof. Since $Z = \prod_k [0, B_k]$ under the assumptions made in Section 2.1, we can decompose Z^L dimension-wise. Define Z^L_k for $k \in [d_{a(n)}]$ as follows:

$$\begin{aligned} Z^L_k = \{(w^+_{nk}, w^-_{nk}) : \exists z_{nk} \in [0, B_k] \text{ and } \exists r_{nk} \in \{0,1\} \text{ s.t.} \\ w^+_{nk} - w^-_{nk} = z_{nk} - \hat{x}_{a(n),k}, \\ 0 \le w^+_{nk} \le (B_k - \hat{x}_{a(n),k}) r_{nk}, \\ 0 \le w^-_{nk} \le \hat{x}_{a(n),k}(1 - r_{nk})\}. \end{aligned}$$

Now, we can write $Z^L = \prod_{k \in [d_{a(n)}]} Z^L_k$. This implies that

$$\text{conv}(Z^L) = \prod_{k \in [d_{a(n)}]} \text{conv}(Z^L_k). \tag{21}$$

This further implies that $\text{relint}(\text{conv}(Z^L)) = \prod_{k \in [d_{a(n)}]} \text{relint}(\text{conv}(Z^L_k))$. Now, it is easy to see that $\text{conv}(Z^L_k)$ is the convex hull of points $(0,0)$, $(B_k - \hat{x}_{nk}, 0)$ and $(0, \hat{x}_{nk})$. The given (u^+_{nk}, u^-_{nk}) in (20) thus belongs to $\text{relint}(\text{conv}(Z^L_k))$. Consequently, the entire vector $(u^+_{nk}, u^-_{nk})_{k \in [d_{a(n)}]}$ defined in (20) belongs to $\text{relint}(\text{conv}(Z^L))$.

Next, we establish the second requirement: $u_{n0} > \overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) - \hat{\theta}_n$. Since $\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)})$ and the closed convex envelope satisfies $\overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) \le \underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})$, we have

$$\begin{aligned} \overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) - \hat{\theta}_n &\le \underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)}) - \hat{\theta}_n \\ &= \underline{Q}_n(\hat{x}_{a(n)} + u_n^+ - u_n^-) - \hat{\theta}_n. \end{aligned}$$

The second equality follows from the definition of $\underline{Q}'_n$ in (11). Since $u_n^+ = u_n^-$ componentwise, we have $\hat{x}_{a(n)} + u_n^+ - u_n^- = \hat{x}_{a(n)}$, so $\underline{Q}_n(\hat{x}_{a(n)} + u_n^+ - u_n^-) = \underline{Q}_n(\hat{x}_{a(n)})$. Therefore, $\overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) - \hat{\theta}_n \le \underline{Q}_n(\hat{x}_{a(n)}) - \hat{\theta}_n$. Since $u_{n0} \in (\underline{Q}_n(\hat{x}_{a(n)}) - \hat{\theta}_n, \infty)$, we conclude that $u_{n0} > \overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) - \hat{\theta}_n$, completing the proof. □

The advantage of Proposition 4 is that it provides a computationally efficient way to obtain a core point when the incumbent solution does not lie on the boundary. Specifically, the scalar u_{n0} can be computed directly from $\underline{Q}_n(\hat{x}_{a(n)})$, which is readily available, without requiring additional subproblem solves. For incumbent solutions that lie on the boundary of the feasible region, we discuss our approach for deriving a core point in the computational results section.

Having established Pareto-optimality of the normalized Lagrangian cuts in the extended space (Proposition 3), we now investigate the implications of this property in the original space. Towards this end, we consider the following definition of Pareto-optimal cuts.

Definition 2 (Pareto-optimal h-cut)**.** *A cut of the form $\theta_n \ge h(x_{a(n)}; \ell^1, \pi^1)$ dominates the cut $\theta_n \ge h(x_{a(n)}; \ell^2, \pi^2)$ if $h(x_{a(n)}; \ell^1, \pi^1) \ge h(x_{a(n)}; \ell^2, \pi^2)$ for all $x_{a(n)} \in X_{a(n)}$, with strict inequality for at least one point in $X_{a(n)}$. A valid h-cut is Pareto-optimal for reference set $\text{epi}_{X_{a(n)}}(\underline{Q}_n)$ if no other valid h-cut dominates it.*

Note that Definition 2 generalizes Definition 1 of Pareto-optimal linear cuts. For a given incumbent $\hat{x}_{a(n)}$, the function h of interest is:

$$h(x_{a(n)};\pi_n^+,\pi_n^-,\pi_{n0}) = \frac{1}{\pi_{n0}}\Bigg(\mathcal{L}_n(\pi_n^+,\pi_n^-,\pi_{n0};\hat{x}_{a(n)}) - \sum_{k\in[d_{a(n)}]}\pi_{nk}^+(x_{a(n),k}-\hat{x}_{a(n),k})^+ - \sum_{k\in[d_{a(n)}]}\pi_{nk}^-(x_{a(n),k}-\hat{x}_{a(n),k})^-\Bigg). \quad (22)$$

The cut $\theta_n \geq h(x_{a(n)};\pi_n^+,\pi_n^-,\pi_{n0})$ is equivalent to the ReLU cut (19) in the original space. This formulation allows us to define Pareto-optimality for all valid ReLU cuts at a given incumbent.

In defining Pareto optimality, it is essential to specify the reference set over which Pareto optimality is attained. In our setting, the reference set of interest is defined as follows:

$$\mathcal{H}_n = \{(x_{a(n)},\theta_n) : x_{a(n)} \in \mathrm{conv}(Z_{a(n)}), \theta_n \geq h(x_{a(n)};\pi_n^+,\pi_n^-,\pi_{n0}), \forall \pi_n^+,\pi_n^- \in \mathbb{R}^{d_{a(n)}}, \pi_{n0} \geq 0\}. \quad (23)$$

Both function h and set $\mathcal{H}_n$ depend on the incumbent $\hat{x}_{a(n)}$, although we omit this dependence from the notation for brevity. The set $\mathcal{H}_n$ can be interpreted as the epigraph generated by adding all ReLU cuts of the form (19) for a fixed incumbent $\hat{x}_{a(n)}$. We depict an example of the set $\mathcal{H}_n$ in Figure 1.

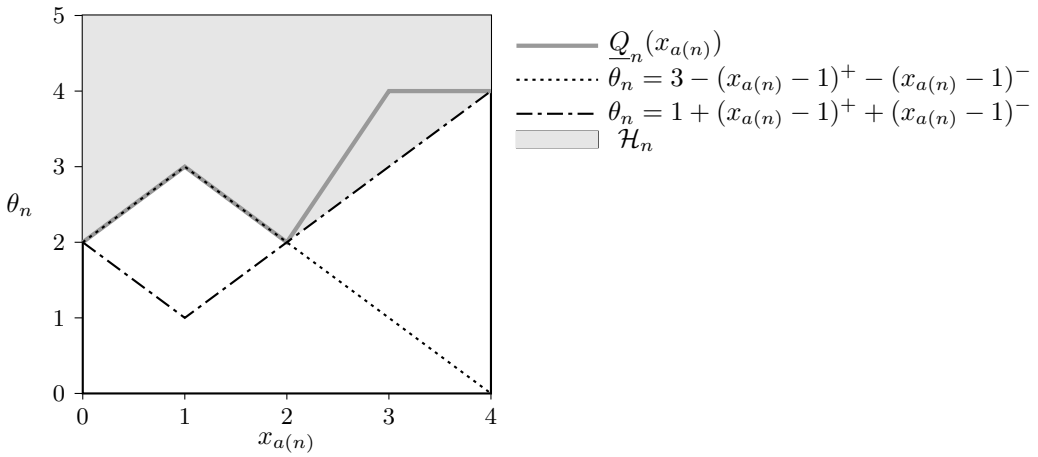

Figure 1: The solid (grey) line depicts $\underline{Q}_n(x_{a(n)})$, a piecewise-linear function. The domain $Z_{a(n)} = [0,4]$ and the shaded region shows the epigraph $\mathcal{H}_n$ at incumbent $\hat{x}_{a(n)} = 1$. Two ReLU cuts are shown: $\theta_n = 3 - (x_{a(n)} - 1)^+ - (x_{a(n)} - 1)^-$ (dotted line) and $\theta_n = 1 + (x_{a(n)} - 1)^+ + (x_{a(n)} - 1)^-$ (dashed line). These cuts are obtained using different normalization coefficients, and their epigraph intersection equals $\mathcal{H}_n$. Observe that $\mathrm{epi}(\underline{Q}_n) \subsetneq \mathcal{H}_n \subsetneq \mathrm{epi}(\overline{\mathrm{co}}(\underline{Q}_n))$.

We now establish a useful property of the proposed cut with respect to this new notion of Pareto optimality.

Proposition 5. *Let $\hat{x}_{a(n)}$ be the incumbent solution and function h be as defined in (22). For all points (u_n^+,u_n^-,u_{n0}) that belong to $\mathrm{relint}\left(\mathrm{epi}(\overline{\mathrm{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))) - (\mathbf{0},\mathbf{0},\hat{\theta}_n)\right)$, any optimal solution $(\hat{\pi}_n^+,\hat{\pi}_n^-,\hat{\pi}_{n0})$ to the normalized dual (17) with $\hat{\pi}_{n0} > 0$ defines a Pareto-optimal h-cut of form*

$\theta_n \geq h(x_{a(n)}; \hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ *for reference set* $\mathcal{H}_n$. *Further, we have* $\mathcal{H}_n \subseteq \text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$, *and if* $\overline{\text{co}}(\underline{Q}_n)(\hat{x}_{a(n)}) < \underline{Q}_n(\hat{x}_{a(n)})$, *then* $\mathcal{H}_n \subsetneq \text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$.

Proof. Consider the dual problem (17) with normalization coefficients

$$(u_n^+, u_n^-, u_{n0}) \in \text{relint}\left(\text{epi}(\overline{\text{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))) - (\mathbf{0}, \mathbf{0}, \hat{\theta}_n)\right).$$

Let $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ be an optimal solution of the dual problem with $\pi_{n0} > 0$. Assume, for a contradiction, that the resulting ReLU cut,

$$\theta_n \geq h(x_{a(n)}; \hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0}) \tag{24}$$

is not a Pareto-optimal h-cut for $\mathcal{H}_n$. This implies that there exists different set of dual variables $(\tilde{\pi}_n^+, \tilde{\pi}_n^-, \tilde{\pi}_{n0})$ with $\tilde{\pi}_{n0} > 0$, such that the cut

$$\theta_n \geq h(x_{a(n)}; \tilde{\pi}_n^+, \tilde{\pi}_n^-, \tilde{\pi}_{n0}) \tag{25}$$

dominates cut (24) for $(x_{a(n)}, \theta_n) \in \mathcal{H}_n$. Now, reformulating the ReLU functions, we can represent cuts (24) and (25) in the form (18), in the extended space $Z^{lift}_{\hat{x}_{a(n)}}$. Specifically, we reformulate (24) as:

$$\theta_n \geq \frac{1}{\hat{\pi}_{n0}}\left(\mathcal{L}_n(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0}; \hat{x}_{a(n)}) - (\hat{\pi}_n^+, \hat{\pi}_n^-)\begin{pmatrix} w_n^+ \\ w_n^- \end{pmatrix}\right). \tag{26}$$

Similarly, we reformulate (25) as:

$$\theta_n \geq \frac{1}{\tilde{\pi}_{n0}}\left(\mathcal{L}_n(\tilde{\pi}_n^+, \tilde{\pi}_n^-, \tilde{\pi}_{n0}; \hat{x}_{a(n)}) - (\tilde{\pi}_n^+, \tilde{\pi}_n^-)\begin{pmatrix} w_n^+ \\ w_n^- \end{pmatrix}\right). \tag{27}$$

We show that if cut (25) dominates (24) on $\mathcal{H}_n$, then cut (27) dominates cut (26) in the extended space $\text{epi}(\overline{\text{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})))$ on $\text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$. This contradicts Proposition 3, which states that cut (26) is Pareto-optimal on $\text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$.

Since, the set $Z_{a(n)} = \prod_{k\in[d_{a(n)}]}[0, B_k]$, we consider points $x_{a(n)} \in \prod_{k\in[d_{a(n)}]}\{0, \hat{x}_{a(n),k}, B_k\} \subset Z_{a(n)}$, where $\prod_{k\in[d_{a(n)}]}\{0, \hat{x}_{a(n),k}, B_k\}$ denotes the cartesian product of the set $\{0, \hat{x}_{a(n),k}, B_k\}$ across all dimensions $k \in [d_{a(n)}]$. Furthermore, since cut (25) dominates cut (24) on $\mathcal{H}_n$, by definition of $\mathcal{H}_n$, the dominance will also hold over $\prod_{k\in[d_{a(n)}]}\{0, \hat{x}_{a(n),k}, B_k\}$. In other words:

$$h(x_{a(n)}; \tilde{\pi}_n^+, \tilde{\pi}_n^-, \tilde{\pi}_{n0}) \geq h(x_{a(n)}; \hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0}), \quad \forall x_{a(n)} \in \prod_{k\in[d_{a(n)}]}\{0, \hat{x}_{a(n),k}, B_k\}. \tag{28}$$

Next, for $i \in [2d_{a(n)}]$, let $e^i \in \mathbb{R}^{2d_{a(n)}}$ denote the i^{th} unit vector, i.e., the vector whose i^{th} component equals 1 and whose remaining components are zero. We index the coordinates of $\mathbb{R}^{2d_{a(n)}}$ so that, for each $k \in [d_{a(n)}]$, the k^{th} component corresponds to the variable w_{nk}^+, while the $(d_{a(n)} + k)^{th}$ component corresponds to the variable w_{nk}^-.

In addition, define scalars v_i for $i \in [2d_{a(n)}]$ by

$$v_k := B_k - \hat{x}_{a(n),k}, \qquad v_{d_{a(n)}+k} := \hat{x}_{a(n),k}, \quad \forall k \in [d_{a(n)}].$$

Now, consider all points $\{v_i e^i\}, i \in [2d_{a(n)}]$. Clearly, they belong to the set $Z^{lift}_{\hat{x}_{a(n)}}$ and, therefore, also to the set $\text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$. Under the mapping (9b), these points belong to $\prod_{k\in[d_{a(n)}]}\{0, \hat{x}_{a(n),k}, B_k\}$ in the $Z_{a(n)}$ space. Furthermore, for these points, the right-hand side of (26) and (27) match with right-hand side of (24) and (25), respectively, on the corresponding points in $\prod_{k\in[d_{a(n)}]}\{0, \hat{x}_{a(n),k}, B_k\}$. Owing to domination relation in (28), this implies that cut (27) dominates cut (26) on points $\{v_i e^i\}, i \in [2d_{a(n)}]$ in the extended space. Now, due to the decomposition relation established in (21), we know that $\text{conv}(\{v_i e^i\}_{i\in[2d_{a(n)}]}) = \text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$. This proves that cut (27) dominates cut (26) on entire $\text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$. This contradicts Proposition 3, hence our original claim in Proposition 5 is true.

Now, we prove the second part of the proposition which states that set $\mathcal{H}_n \subseteq \text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$ and if $\overline{\text{co}}(\underline{Q}_n)(\hat{x}_{a(n)}) < \underline{Q}_n(\hat{x}_{a(n)})$, then $\mathcal{H}_n \subsetneq \text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$. According to Proposition 4 of Deng and Xie 2024, any tight Lagrangian cut derived in the original space, from the Lagrangian dual obtained by relaxing the standard copy constraints, is a ReLU Lagrangian cut (8) with $\mathcal{L}^R_n(\pi^+_n, \pi^-_n; \hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)})$. More generally, the same proof works in showing that any normalized Lagrangian cut derived in the original space can be expressed as a normalized ReLU Lagrangian cut. This implies that the set $\mathcal{H}_n$ defined in (23) also consists of normalized Lagrangian cuts (linear cuts obtained in original space). From Theorem 3.5, Remark 3.13, and Lemma 3.14 of Füllner et al. 2024b, the collection of all normalized Lagrangian cuts recovers the epigraph $\text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$. It therefore follows that $\mathcal{H}_n$ with linear Lagrangian cuts and additional non-linear ReLU cuts is a subset of $\text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$.

Next consider the case when $\overline{\text{co}}(\underline{Q}_n)(\hat{x}_{a(n)}) < \underline{Q}_n(\hat{x}_{a(n)})$. We know that point $(\hat{x}_{a(n)}, \overline{\text{co}}(\underline{Q}_n)(\hat{x}_{a(n)}))$ belongs to $\text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$. Since, $\overline{\text{co}}(\underline{Q}_n)(\hat{x}_{a(n)}) < \underline{Q}_n(\hat{x}_{a(n)})$, we also know that there exists a ReLU cut that violates this point. So, this point does not belong to $\mathcal{H}_n$. This proves that $\mathcal{H}_n \subsetneq \text{epi}_{\text{conv}(Z_{a(n)})}(\overline{\text{co}}(\underline{Q}_n))$. □

Note that if we do not employ ReLU copy constraints and instead use the standard copy constraints, then normalizing the respective dual yields linear Lagrangian cuts. As shown by Füllner et al. 2024b, such cuts can attain Pareto-optimality only with respect to the epigraph $\text{epi}_{\text{conv}(Z_{a(n)})}\left(\overline{\text{co}}\left(\underline{Q}_n\right)\right)$. Since the set $\mathcal{H}_n$ is a strict subset of this epigraph, Proposition 5 establishes that the nonlinear ReLU cuts achieve a stronger notion of Pareto-optimality than the linear cuts derived by normalizing the standard Lagrangian dual.

2.4.2 Tight Cuts

In discussing tight cuts, it is important to distinguish between two notions of cut strength that are often conflated in the literature. The first is tightness at incumbent solution: a cut is tight if it matches the cost-to-go function $\underline{Q}_n$ exactly at the current state variable solution. Under this definition, several families of cuts—including ReLU cuts obtained by solving dual (7)—are tight. For instance, Λ-shaped cuts (see Ahmed et al. 2022, Deng and Xie 2024), which extend classical L-shaped cuts to mixed-integer state variables and are obtained by imposing $\pi^+_n = \pi^-_n$ in (7), are also tight at the incumbent. Tightness is a convenient sufficient condition for convergence—if one generates a tight cut at every iteration, asymptotic convergence follows. It is not, however, a necessary condition. Proposition 2 establishes asymptotic convergence without requiring tightness at the incumbent.

The second notion is global approximation quality: a cut is globally strong if it yields strong lower

bounds not only at the incumbent but also across the feasible region. Λ-shaped cuts are typically weak in this sense, since their approximation can be weak away from the incumbent (see computational studies of integer L-shaped cuts (Zou et al. 2019, Bansal and Küçükyavuz 2024) and of Λ-shaped cuts (Deng and Xie 2024) for reference). For this reason, the Pareto-optimality concept introduced in the previous section provides a more meaningful measure of approximation quality. In this section, we study how to select normalization coefficients to obtain ReLU cuts that are tight at the incumbent while also being Pareto-optimal over the domain.

Proposition 6. *There exists an ϵ-ball $B_\epsilon(\mathbf{0},\mathbf{0})$, around $(\mathbf{0},\mathbf{0})$ in extended space $\operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}})$ such that for normalization coefficients (u_n^+, u_n^-, u_{n0}) satisfying:*

$$(u_n^+, u_n^-) \in B_\epsilon(\mathbf{0},\mathbf{0}) \cap \operatorname{relint}(\operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}})),$$
$$u_{n0} \geq \underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)}) - \hat{\theta}_n,$$

an optimal dual solution $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ to (17) with $\hat{\pi}_{n0} > 0$ defines a tight ReLU cut at the incumbent solution $\hat{x}_{a(n)}$, that is, $\frac{1}{\hat{\pi}_{n0}}\big(\mathcal{L}_n(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0}; \hat{x}_{a(n)})\big) = \underline{Q}_n(\hat{x}_{a(n)})$. Furthermore, the cut is a Pareto-optimal h-cut over the reference set $\mathcal{H}_n$ for h defined in (22).

Proof. We prove this result in four parts.

6.1 We first show that the function $\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})$ is piecewise polyhedral with finitely many pieces.

6.2 There is a neighborhood $B_\epsilon(\mathbf{0},\mathbf{0})$ of the extreme point $(\mathbf{0},\mathbf{0})$ in the lifted space $\operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}})$ where all affine pieces that locally define the function $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$ must agree at $(\mathbf{0},\mathbf{0})$. Furthermore, the value of the closed convex envelope at $(\mathbf{0},\mathbf{0})$ is $\underline{Q}'_n((\mathbf{0},\mathbf{0});\hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)})$.

6.3 We show that if the coefficients (u_n^+, u_n^-, u_{n0}) satisfy the conditions of the proposition, then an optimal solution of (17) induces a cut of the form (18) that is tight at some point $(\tilde{w}_n^+, \tilde{w}_n^-) \in \operatorname{relint}(B_\epsilon(\mathbf{0},\mathbf{0}))$ with respect to the closed convex envelope $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$.

6.4 If the cut is tight at $(\tilde{w}_n^+, \tilde{w}_n^-) \in \operatorname{relint}(B_\epsilon(\mathbf{0},\mathbf{0}))$ for $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$, then it is also tight at $(\mathbf{0},\mathbf{0})$ for $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$. By Part 6.2, tightness at $(\mathbf{0},\mathbf{0})$ for $\overline{\operatorname{co}}(\underline{Q}'_n)$ implies tightness at $(\mathbf{0},\mathbf{0})$ with respect to $\underline{Q}'_n$. Projecting to the original space, this yields tightness with respect to $\underline{Q}_n$ at $\hat{x}_{a(n)}$.. Furthermore, the cut written in the original space is a Pareto-optimal h-cut over the reference set $\mathcal{H}_n$ for h defined in (22).

We first prove **Part** 6.1. By Lemma 2.2 of Füllner et al. 2024b, the function $\underline{Q}_n$ is piecewise polyhedral with finitely many pieces. Observe that the function $\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)})$ can be written as the composition $\underline{Q}_n \circ G$, where the affine mapping $G : \mathbb{R}^{2d_{a(n)}} \to \mathbb{R}^{d_{a(n)}}$ is defined as $G(w_n^+, w_n^-) = \hat{x}_{a(n)} + w_n^+ - w_n^-$. For notational simplicity, we omit the dependence of the function G on the incumbent $\hat{x}_{a(n)}$. To show that $\underline{Q}_n \circ G$ is piecewise polyhedral, it suffices to prove that its epigraph is a finite union of polyhedra. Indeed,

$$\begin{aligned}\operatorname{epi}(\underline{Q}_n \circ G) &= \Big\{(w_n^+, w_n^-, \theta_n) : \theta_n \geq \underline{Q}_n\big(G(w_n^+, w_n^-)\big)\Big\} \\ &= \Big\{(w_n^+, w_n^-, \theta_n) : \big(G(w_n^+, w_n^-), \theta_n\big) \in \operatorname{epi}(\underline{Q}_n)\Big\}.\end{aligned}$$

Define the affine map $V : \mathbb{R}^{2d_{a(n)}+1} \to \mathbb{R}^{d_{a(n)}+1}$ as $V(w_n^+, w_n^-, \theta_n) = \big(G(w_n^+, w_n^-), \theta_n\big)$. With this notation, $\operatorname{epi}(\underline{Q}_n \circ G) = V^{-1}\big(\operatorname{epi}(\underline{Q}_n)\big)$. Since $\underline{Q}_n$ is piecewise polyhedral, its epigraph $\operatorname{epi}(\underline{Q}_n)$

can be expressed as a finite union of polyhedra. The preimage of a polyhedron under an affine map is again a polyhedron. Consequently, $V^{-1}\big(\operatorname{epi}(\underline{Q}_n)\big)$ is a finite union of polyhedra and hence $\operatorname{epi}(\underline{Q}_n \circ G)$ is piecewise polyhedral. This completes the proof of the first part.

We next prove **Part** 6.2. Since $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$ is a polyhedral convex function, it admits a finite max-representation. The representation is finite because of Part 6.1. Specifically, there exists a finite index set I and affine functions

$$\ell_i(w_n^+, w_n^-) = \alpha_i^\top w_n^+ + \beta_i^\top w_n^- + \gamma_i, \qquad i \in I, \tag{29}$$

such that $\overline{\operatorname{co}}\big(\underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)})\big) = \max_{i\in I} \ell_i(w_n^+, w_n^-)$. Now, define

$$M := \overline{\operatorname{co}}\big(\underline{Q}'_n(\mathbf{0},\mathbf{0}; \hat{x}_{a(n)})\big) = \max_{i\in I} \ell_i(\mathbf{0},\mathbf{0}) = \max_{i\in I} \gamma_i,$$

where the second equality follows from substituting $(\mathbf{0},\mathbf{0})$ in (29). Let $I^\star := \{i \in I : \ell_i(\mathbf{0},\mathbf{0}) = M\}$ denote the set of affine pieces attaining the maximum at $(\mathbf{0},\mathbf{0})$. If $I^\star = I$, then any $\epsilon > 0$ suffices for the ball $B_\epsilon(\mathbf{0},\mathbf{0})$. Otherwise, we construct ϵ as follows. Fix any $i \notin I^\star$ (such an i exists since $I^\star \subsetneq I$). Then $\ell_i(\mathbf{0},\mathbf{0}) < M$. Choose any $j \in I^\star$ and consider the affine difference

$$d_{ij}(w_n^+, w_n^-) := \ell_j(w_n^+, w_n^-) - \ell_i(w_n^+, w_n^-).$$

Since d_{ij} is affine and

$$\begin{aligned} d_{ij}(\mathbf{0},\mathbf{0}) &= \ell_j(\mathbf{0},\mathbf{0}) - \ell_i(\mathbf{0},\mathbf{0}) \\ &= M - \ell_i(\mathbf{0},\mathbf{0}) > 0, \end{aligned}$$

by continuity, there exists $\epsilon_{ij} > 0$ such that

$$d_{ij}(w_n^+, w_n^-) > 0 \quad \text{for all } (w_n^+, w_n^-) \in B_{\epsilon_{ij}}(\mathbf{0},\mathbf{0}).$$

Equivalently,

$$\ell_i(w_n^+, w_n^-) < \ell_j(w_n^+, w_n^-) \quad \text{for all } (w_n^+, w_n^-) \in B_{\epsilon_{ij}}(\mathbf{0},\mathbf{0}).$$

Now define $\epsilon_i := \min_{j\in I^\star} \epsilon_{ij} > 0$. Then, for all $(w_n^+, w_n^-) \in B_{\epsilon_i}(\mathbf{0},\mathbf{0})$,

$$\begin{aligned} \ell_i(w_n^+, w_n^-) &< \ell_j(w_n^+, w_n^-), \qquad \forall j \in I^\star, \\ \ell_i(w_n^+, w_n^-) &< \max_{j\in I^\star} \ell_j(w_n^+, w_n^-) \\ &\le \max_{m\in I} \ell_m(w_n^+, w_n^-) \\ &= \overline{\operatorname{co}}\big(\underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)})\big). \end{aligned}$$

Thus, no affine piece ℓ_i with $i \notin I^\star$ can locally define the function in $B_{\epsilon_i}(\mathbf{0},\mathbf{0})$. Since there are finitely many indices $i \notin I^\star$, let $\epsilon := \min_{i\notin I^\star} \epsilon_i > 0$. Then, for every $(w_n^+, w_n^-) \in B_\epsilon(\mathbf{0},\mathbf{0})$, any affine piece that locally defines $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot,\cdot;\hat{x}_{a(n)}))$ must belong to $I^\star$ and therefore agree at $(\mathbf{0},\mathbf{0})$. This proves the first statement in Part 6.2. The second statement in this part follows immediately from the proof of Proposition 2.

Now, we prove **Part** 6.3. We use Lemma 3.21 of Füllner et al. 2024b, which states that the normalized

dual problem (17) can be formulated as an LP, and the dual of the LP is given by:

$$\begin{aligned} \min_{\lambda_n, w_n^+, w_n^-, \eta_n} \quad & \eta_n \\ & (\lambda_n, w_n^+, w_n^-) \in \operatorname{conv}(\mathcal{W}_{\hat{x}_{a(n)}}) \\ & \eta_n \geq 0, \\ & u_{n0}\eta_n \geq c_n^\top \lambda_n - \hat{\theta}_n, \\ & \eta_n \begin{pmatrix} u_n^+ \\ u_n^- \end{pmatrix} = \begin{pmatrix} w_n^+ \\ w_n^- \end{pmatrix} - \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}. \end{aligned} \tag{30}$$

Here, the notation λ_n denotes the set of variables $(x_n, y_n, (\theta_m)_{m\in C(n)})$, and $c_n^\top \lambda_n$ denotes the objective function $f_n(x_n, y_n) + \sum_{m\in C(n)} q_{nm}\theta_m$. Further, the set $\mathcal{W}_{\hat{x}_{a(n)}}$ is defined as:

$$\begin{aligned} \mathcal{W}_{\hat{x}_{a(n)}} := \{(\lambda_n, w_n^+, w_n^-) : & (x_n, y_n) \in H_n(\hat{x}_{a(n)} + w_n^+ - w_n^-) \cap (X_n \times Y_n), \\ & (w_n^+, w_n^-) \in Z^{lift}_{\hat{x}_{a(n)}}, \\ & (x_n, \theta_m) \in \Psi_m, \quad \forall m \in C(n)\}. \end{aligned}$$

Now consider the ball $B_\epsilon(\mathbf{0}, \mathbf{0})$ defined in the proof of Part 6.2. The relatively complete recourse assumption $\operatorname{dom}(\underline{Q}_n) = Z_{a(n)} = \prod_k [0, B_k]$ ensures that for every $(w_n^+, w_n^-) \in \operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}})$, the corresponding state $\hat{x}_{a(n)} + w_n^+ - w_n^-$ lies in the domain of $\underline{Q}_n$, and hence the feasible set $H_n(\hat{x}_{a(n)} + w_n^+ - w_n^-)$ is non-empty. Consequently, for $(w_n^+, w_n^-) \in \operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}})$, there exists $\lambda_n = (x_n, y_n, (\theta_m)_{m\in C(n)})$ such that $(\lambda_n, w_n^+, w_n^-) \in \mathcal{W}_{\hat{x}_{a(n)}}$. In other words,

$$B_\epsilon(\mathbf{0}, \mathbf{0}) \subseteq \operatorname{Proj}_{w_n^+, w_n^-} \operatorname{conv}(\mathcal{W}_{\hat{x}_{a(n)}}).$$

Now fix any $(u_n^+, u_n^-) \in \operatorname{relint}(B_\epsilon(\mathbf{0}, \mathbf{0})) \cap \operatorname{relint}(\operatorname{conv}(Z^{lift}_{\hat{x}_{a(n)}}))$ and choose $u_{n0} \geq \underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)}) - \hat{\theta}_n$. We claim that the LP dual (30) has optimal value $\eta^* \leq 1$. To see this, consider the candidate solution $\eta_n = 1$ and $(w_n^+, w_n^-) = (u_n^+, u_n^-)$. By the inclusion above, and definition of $\mathcal{W}_{\hat{x}_{a(n)}}$ and $\underline{Q}'_n$, there exists λ_n such that $(\lambda_n, u_n^+, u_n^-) \in \operatorname{conv}(\mathcal{W}_{\hat{x}_{a(n)}})$ and $c_n^\top \lambda_n \leq \underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})$. The constraint $\eta_n(u_n^+, u_n^-)^\top = (w_n^+, w_n^-)^\top - (\mathbf{0}, \mathbf{0})^\top$ is satisfied since $\eta_n = 1$. Further, since $u_{n0} \geq \underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)}) - \hat{\theta}_n$, the constraint $u_{n0}\eta_n \geq c_n^\top \lambda_n - \hat{\theta}_n$ is also satisfied. Hence $\eta_n = 1$ is feasible, implying $\eta^* \leq 1$.

By Corollary 3.22 of Füllner et al. 2024b, the projection of $((\mathbf{0}, \mathbf{0}), \hat{\theta}_n)$ onto $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)})))$ along direction (u_n^+, u_n^-, u_{n0}) is given by

$$(\tilde{w}_n^+, \tilde{w}_n^-, \tilde{\theta}_n) = (\mathbf{0}, \mathbf{0}, \hat{\theta}_n) + \eta_n^*(u_n^+, u_n^-, u_{n0}), \tag{31}$$

and the normalized cut (18) supports $\operatorname{epi}(\overline{\operatorname{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)})))$ at this point. Since $\eta_n^* \leq 1$ and $(u_n^+, u_n^-) \in \operatorname{relint}(B_\epsilon(\mathbf{0}, \mathbf{0}))$, we have $(\tilde{w}_n^+, \tilde{w}_n^-) = \eta_n^*(u_n^+, u_n^-) \in \operatorname{relint}(B_\epsilon(\mathbf{0}, \mathbf{0}))$. This completes the proof of Part 6.3.

Finally, we prove **Part** 6.4. Recall from Part 6.2, that

$$\ell_j(\mathbf{0}, \mathbf{0}) = \gamma_j = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)}), \text{ for all } j \in I^\star.$$

For simplicity, we denote the value $\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$ as Q_0 in rest of this proof. Now, the function $\overline{\operatorname{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$ for $(w_n^+, w_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0})$ can be written as:

$$\overline{\operatorname{co}}(\underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)})) = Q_0 + \max_{j\in I^\star} \alpha_j^\top w_n^+ + \beta_j^\top w_n^- =: Q_0 + p(w_n^+, w_n^-),$$

where $p(w_n^+, w_n^-) := \max_{j \in I^\star} \alpha_j^\top w_n^+ + \beta_j^\top w_n^-$. Note that the function p is positively homogeneous, that is, $p(tw_n^+, tw_n^-) = tp(w_n^+, w_n^-)$ for $t \geq 0$. Now, consider the normalized cut obtained in Part 6.3. Again, for simplicity, we denote this cut as $\theta_n \geq c^+ w_n^+ + c^- w_n^- + c_0$. Since the cut is valid,

$$c^+ w_n^+ + c^- w_n^- + c_0 \leq \overline{\text{co}}(\underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)})), \quad \forall (w_n^+, w_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0}).$$

This implies that

$$c^+ w_n^+ + c^- w_n^- + c_0 - Q_0 \leq p(w_n^+, w_n^-), \quad \forall (w_n^+, w_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0}).$$

Now, for $(\tilde{w}_n^+, \tilde{w}_n^-)$ identified in Part 6.3, we have $c^+ \tilde{w}_n^+ + c^- \tilde{w}_n^- + c_0 = \overline{\text{co}}(\underline{Q}'_n(\tilde{w}_n^+, \tilde{w}_n^-; \hat{x}_{a(n)}))$, therefore, the above implication leads to

$$c^+ \tilde{w}_n^+ + c^- \tilde{w}_n^- + c_0 - Q_0 = p(\tilde{w}_n^+, \tilde{w}_n^-).$$

Furthermore, since $(\tilde{w}_n^+, \tilde{w}_n^-) \in \text{relint}(B_\epsilon(\mathbf{0}, \mathbf{0}))$ and also $(\mathbf{0}, \mathbf{0}) \in B_\epsilon(\mathbf{0}, \mathbf{0})$, there exists $\delta > 0$ such that $t(\tilde{w}_n^+, \tilde{w}_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0})$ for all $t \in [1 - \delta, 1 + \delta]$. For such t, we have:

$$t(c^+ \tilde{w}_n^+ + c^- \tilde{w}_n^-) + c_0 - Q_0 \leq p(t\tilde{w}_n^+, t\tilde{w}_n^-) = tp(\tilde{w}_n^+, \tilde{w}_n^-) = t(c^+ \tilde{w}_n^+ + c^- \tilde{w}_n^- + c_0 - Q_0).$$

This gives: $c_0 - Q_0 \leq t(c_0 - Q_0)$, which implies that $(c_0 - Q_0)(1-t) \leq 0$ for all $t \in [1-\delta, 1+\delta]$. Taking $t < 1$ gives $c_0 - Q_0 \leq 0$ and $t > 1$ gives $c_0 - Q_0 \geq 0$, and hence $c_0 - Q_0 = 0$ or $c_0 = Q_0$. This proves that the obtained cut of the form $\theta_n \geq c^+ w_n^+ + c^- w_n^- + c_0$ is tight at $(\mathbf{0}, \mathbf{0})$ for $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$. By Part 6.2, tightness at $(\mathbf{0}, \mathbf{0})$ for $\overline{\text{co}}(\underline{Q}'_n)$ implies tightness at $(\mathbf{0}, \mathbf{0})$ with respect to $\underline{Q}'_n$. Projecting to the original space, this yields tightness with respect to $\underline{Q}_n$ at $\hat{x}_{a(n)}$. Moreover, the coefficients (u_n^+, u_n^-, u_{n0}) selected in Part 6.3 satisfy the requirements of Proposition 5, so the normalized cut is also a Pareto-optimal h-cut. □

The main insight of Proposition 6 is that tight cuts can be attained through an appropriate choice of the core point. In particular, the proposition suggests selecting the core point sufficiently close to $(\mathbf{0}, \mathbf{0})$ in order to obtain a cut that is tight at the incumbent. We next strengthen Proposition 6 by showing that for any vectors (u_n^+, u_n^-) satisfying the relative-interior requirement, we can tune the scalar u_{n0} to produce a tight cut.

Proposition 7. *There exists a scalar $\alpha > 1$ such that for normalization coefficients (u_n^+, u_n^-, u_{n0}) satisfying:*

$$(u_n^+, u_n^-) \in \text{relint}(\text{conv}(Z^{lift}_{\hat{x}_{a(n)}}))$$
$$u_{n0} \geq \alpha \left(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)}) - \hat{\theta}_n \right),$$

the optimal dual solution $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ to (17) with $\hat{\pi}_{n0} > 0$ defines a tight ReLU cut at incumbent solution $\hat{x}_{a(n)}$. Furthermore, the cut is Pareto-optimal h-cut over the reference set $\mathcal{H}_n$ for h defined in (22).

The proof of Proposition 7 is provided in Appendix B. The proposition shows that, for any core point, one can obtain a tight cut by choosing the coefficient u_{n0} sufficiently large. Together, Propositions 6 and 7 establish normalization as a method to construct cuts that are simultaneously tight and Pareto-optimal. Moreover, our arguments extend to the setting with standard copy constraints (as opposed to ReLU-based constraints), in which case the resulting cuts reduce to linear normalized Lagrangian cuts. Recent literature (e.g., Yang and Yang 2025) also proposes an alternative approach to producing tight, Pareto-optimal cuts; we contrast this approach with our method in the next section.

3 Normalization vs. Regularization

We first revisit recently proposed regularization-based approaches for constructing strong cuts when dual degeneracy can yield cuts with poor approximation quality. We then relate these approaches to normalization and highlight the advantages of the normalization perspective.

The key idea behind regularization is to characterize (or implicitly represent) the set of all optimal solutions to dual (7), and then select a desired solution by optimizing a regularized objective over this set. This objective augments the original dual objective with additional terms involving the dual variables, weighted by appropriately chosen coefficients that ensure tight, Pareto-optimal cuts. We refer to this approach as regularization in analogy with machine learning, where one augments a loss function with penalty terms to bias the optimizer toward solutions with preferred structure.

Formally, let $\Pi_n(\hat{x}_{a(n)})$ denote the set of all optimal solutions of the ReLU dual (7). Given a core point $(\tilde{u}_n^+, \tilde{u}_n^-) \in \text{relint}(\text{conv}(Z^{lift}_{\hat{x}_{a(n)}}))$, Yang and Yang 2025 prove that the cut coefficients resulting from solving the following problem:

$$\max_{\pi_n^+, \pi_n^- \in \Pi_n(\hat{x}_{a(n)})} \mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) - (\pi_n^+, \pi_n^-) \begin{pmatrix} \tilde{u}_n^+ \\ \tilde{u}_n^- \end{pmatrix}, \tag{32}$$

produces a tight and Pareto-optimal cut of the form (8). This approach is inspired by the classical work of Magnanti and Wong 1981 on accelerating Benders decomposition when LP subproblems have dual degeneracy. Yang and Yang 2025 show that the set $\Pi_n(\hat{x}_{a(n)})$ can be modeled using the constraint

$$\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) \geq \underline{Q}_n(\hat{x}_{a(n)}) - \epsilon \tag{33}$$

where $\epsilon > 0$ is a small value. Since the function $\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)})$ is concave and piecewise linear, it is approximated iteratively using gradient cuts via the level-bundle method (Lemaréchal et al. 1995). The normalized dual problem (17) is also solved using the level-bundle method. Moreover, note that the Lagrangian relaxation $\mathcal{L}_n$ in the extended space is simply a mixed-integer reformulation of the Lagrangian relaxation $\mathcal{L}_n^R$ in the original space. Consequently, the computational complexity of solving the regularized and normalized duals is comparable.

Deng and Xie 2024 follow a similar approach to (32) but instead of approximating the set $\Pi_n(\hat{x}_{a(n)})$ exactly using convex program of the form (33), they approximate it using an LP. Specifically, instead of using the function $\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)})$ defined in (6), they use the LP relaxation of $\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)})$ with additional constraints to prevent the approximate linear program from becoming infeasible. Furthermore, the authors discuss only the choice of regularization coefficients that yield a bounded linear program, but provide no guarantees regarding cut quality. The motivation for this LP-based approach is to recover extreme points of $\Pi_n(\hat{x}_{a(n)})$, which correspond to facet-defining cuts in the extended space. However, because the LP only approximates the true optimal set, it may introduce spurious extreme points or exclude existing ones, thereby producing weaker cuts. We empirically compare our normalization approach with both regularization-based methods in the computational results section.

Next, we show that any cut attained by the regularization method (32) can also be attained by solving the normalized dual problem (17) using appropriate normalization coefficients.

Proposition 8. *For $(\tilde{u}_n^+, \tilde{u}_n^-) \in \text{relint}(\text{conv}(Z^{lift}_{\hat{x}_{a(n)}}))$, consider a cut of the form (8) obtained by solving the regularized dual (32). Then, there exist normalization coefficients $(\hat{u}_n^+, \hat{u}_n^-, \hat{u}_{n0})$ such that the coefficients of the regularized cut (up to scaling) are also optimal to the normalized dual (17).*

Proof. We prove the result in the following steps:

8.1 For a given core point $(\tilde{u}_n^+, \tilde{u}_n^-)$, consider the regularized dual in (32). Let $(\tilde{\pi}_n^+, \tilde{\pi}_n^-)$ be any optimal solution of this dual problem. We first show that the resulting cut of the form:

$$\theta_n \geq \mathcal{L}_n^R(\tilde{\pi}_n^+, \tilde{\pi}_n^-; \hat{x}_{a(n)}) - (\tilde{\pi}_n^+, \tilde{\pi}_n^-) \cdot \begin{pmatrix} w_n^+ \\ w_n^- \end{pmatrix}, \tag{34}$$

supports $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$ on the segment $\mathcal{S} = \{t(\tilde{u}_n^+, \tilde{u}_n^-) : t \in [0, \epsilon']\}$, for some $\epsilon' > 0$.

Since $(\tilde{\pi}_n^+, \tilde{\pi}_n^-) \in \Pi_n(\hat{x}_{a(n)})$, we have by (strong) duality that

$$\mathcal{L}_n^R(\tilde{\pi}_n^+, \tilde{\pi}_n^-; \hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)}) = \overline{\text{co}}(\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})),$$

and therefore (34) passes through $\big((\mathbf{0}, \mathbf{0}), \overline{\text{co}}(\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)}))\big)$ in the lifted space.

Moreover, by Part 6.2 of Proposition 6, there exist $\epsilon > 0$, a finite index set $I^\star$, and vectors $\{(\alpha_j, \beta_j)\}_{j \in I^\star}$ such that, for all $(w_n^+, w_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0}) \cap \text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$,

$$\overline{\text{co}}(\underline{Q}'_n(w_n^+, w_n^-; \hat{x}_{a(n)})) = \underline{Q}_n(\hat{x}_{a(n)}) + \max_{j \in I^\star} \left\{ \alpha_j^\top w_n^+ + \beta_j^\top w_n^- \right\}. \tag{35}$$

In particular, every affine function on the right-hand side of (35) is a supporting hyperplane of $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$ that is tight at $(\mathbf{0}, \mathbf{0})$. Since (32) maximizes a linear functional over the set $\Pi_n(\hat{x}_{a(n)})$ of dual-optimal cut coefficients (all of which are tight at $(\mathbf{0}, \mathbf{0})$), the optimality of $(\tilde{\pi}_n^+, \tilde{\pi}_n^-)$ implies that

$$(-\tilde{\pi}_n^+, -\tilde{\pi}_n^-) \cdot \begin{pmatrix} \tilde{u}_n^+ \\ \tilde{u}_n^- \end{pmatrix} = \max_{j \in I^\star} \left\{ (\alpha_j, \beta_j) \cdot \begin{pmatrix} \tilde{u}_n^+ \\ \tilde{u}_n^- \end{pmatrix} \right\} =: \kappa.$$

Choose

$$\epsilon' \;:=\; \min \left\{ 1, \; \frac{\epsilon}{\left\| (\tilde{u}_n^+, \tilde{u}_n^-) \right\|_2} \right\},$$

so that $t(\tilde{u}_n^+, \tilde{u}_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0}) \cap \text{conv}(Z^{lift}_{\hat{x}_{a(n)}})$ for all $t \in [0, \epsilon']$. Then for any $t \in [0, \epsilon']$, using (35) and the definition of κ, we obtain

$$\begin{aligned} \overline{\text{co}}\big(\underline{Q}'_n(t\,\tilde{u}_n^+, t\,\tilde{u}_n^-; \hat{x}_{a(n)})\big) &= \underline{Q}_n(\hat{x}_{a(n)}) + \max_{j \in I^\star} \left\{ t\,(\alpha_j, \beta_j) \cdot \begin{pmatrix} \tilde{u}_n^+ \\ \tilde{u}_n^- \end{pmatrix} \right\} \\ &= \underline{Q}_n(\hat{x}_{a(n)}) + t\,\kappa \\ &= \mathcal{L}_n^R(\tilde{\pi}_n^+, \tilde{\pi}_n^-; \hat{x}_{a(n)}) - (\tilde{\pi}_n^+, \tilde{\pi}_n^-) \cdot \begin{pmatrix} t\,\tilde{u}_n^+ \\ t\,\tilde{u}_n^- \end{pmatrix}. \end{aligned}$$

Hence, (34) is tight (and therefore supports $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$) at every point of the segment $\mathcal{S}$.

8.2 Next, we consider a point $(\hat{u}_n^+, \hat{u}_n^-) \in \mathcal{S} \cap B_\epsilon(\mathbf{0}, \mathbf{0})$ where the ball $B_\epsilon(\mathbf{0}, \mathbf{0})$ is defined in Proposition 6. Such a point always exists because by choosing t sufficiently small, we can ensure that $t(\tilde{u}_n^+, \tilde{u}_n^-) \in B_\epsilon(\mathbf{0}, \mathbf{0})$. The scalar $\hat{u}_{n0}$ is then chosen to satisfy the requirements in Proposition 6, that is, $\hat{u}_{n0} = \underline{Q}'_n(\hat{u}_n^+, \hat{u}_n^-; \hat{x}_{a(n)}) - \hat{\theta}_n$. Now, by Corollary 3.22 of Füllner et al. 2024b (see also discussion around (31)), the normalized dual (17) with normalization coefficients $(\hat{u}_n^+, \hat{u}_n^-, \hat{u}_{n0})$ produces a cut of the form (18) that supports $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$ at a point $\eta(\hat{u}_n^+, \hat{u}_n^-)$ for some $0 < \eta \leq 1$ (and $(\mathbf{0}, \mathbf{0}) \in \mathcal{S}$).

8.3 Let the optimal solution of the normalized dual be $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$. We know that the resulting cut is tight at $(\mathbf{0}, \mathbf{0})$; this follows from Proposition 6. This implies that the following equality holds,

$$\hat{\pi}_{n0}\underline{Q}_n(\hat{x}_{a(n)}) = \mathcal{L}_n(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0}; \hat{x}_{a(n)}).$$

Therefore, the optimal objective value in the normalized dual is $\hat{\pi}_{n0}(\underline{Q}_n(\hat{x}_{a(n)}) - \hat{\theta}_n)$ for the solution $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$.

8.4 From Part 8.2, the normalized cut supports $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$ at $\eta(\hat{u}_n^+, \hat{u}_n^-)$, which means that:

$$\underline{Q}_n(\hat{x}_{a(n)}) - \frac{\eta}{\hat{\pi}_{n0}}(\hat{\pi}_n^+, \hat{\pi}_n^-) \cdot \begin{pmatrix} \hat{u}_n^+ \\ \hat{u}_n^- \end{pmatrix} = \overline{\text{co}}\big(\underline{Q}'_n(\eta\hat{u}_n^+, \eta\hat{u}_n^-; \hat{x}_{a(n)})\big).$$

8.5 From Part 8.1, the regularized cut (34) also supports $\overline{\text{co}}(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)}))$ at $\eta(\hat{u}_n^+, \hat{u}_n^-)$, which implies that

$$\underline{Q}_n(\hat{x}_{a(n)}) - \eta(\tilde{\pi}_n^+, \tilde{\pi}_n^-) \cdot \begin{pmatrix} \hat{u}_n^+ \\ \hat{u}_n^- \end{pmatrix} = \overline{\text{co}}\big(\underline{Q}'_n(\eta\hat{u}_n^+, \eta\hat{u}_n^-; \hat{x}_{a(n)})\big).$$

8.6 Now, we prove that some scaling of the regularized cut is also optimal to the normalized dual. In particular, the solution of interest is $(\tilde{\pi}_n^+, \tilde{\pi}_n^-, 1)$. Note, that the regularized cut (9a) with coefficients $(\tilde{\pi}_n^+, \tilde{\pi}_n^-)$ is the same as the normalized cut (19) with coefficients $(\tilde{\pi}_n^+, \tilde{\pi}_n^-, 1)$. Now, we consider the following scaled solution $(\hat{\pi}_{n0}\tilde{\pi}_n^+, \hat{\pi}_{n0}\tilde{\pi}_n^-, \hat{\pi}_{n0})$, and on substituting in function g_n defining the normalization constraint, we have

$$\begin{aligned} g_n(\hat{\pi}_{n0}\tilde{\pi}_n^+, \hat{\pi}_{n0}\tilde{\pi}_n^-, \hat{\pi}_{n0}) &= \hat{\pi}_{n0}\left((\tilde{\pi}_n^+, \tilde{\pi}_n^-) \cdot \begin{pmatrix} \hat{u}_n^+ \\ \hat{u}_n^- \end{pmatrix} + \hat{u}_{n0}\right) \\ &= \hat{\pi}_{n0}\frac{1}{\eta}\left(\underline{Q}_n(\hat{x}_{a(n)}) - \overline{\text{co}}\big(\underline{Q}'_n(\eta\hat{u}_n^+, \eta\hat{u}_n^-; \hat{x}_{a(n)})\big)\right) + \hat{\pi}_{n0}\hat{u}_{n0} \\ &= (\hat{\pi}_n^+, \hat{\pi}_n^-) \cdot \begin{pmatrix} \hat{u}_n^+ \\ \hat{u}_n^- \end{pmatrix} + \hat{\pi}_{n0}\hat{u}_{n0} \leq 1. \end{aligned}$$

The second equality follows from Part 8.5 and the third equality follows from Part 8.4. The final inequality follows from the fact that $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ is a feasible solution to the normalized dual problem. This means that the solution $(\hat{\pi}_{n0}\tilde{\pi}_n^+, \hat{\pi}_{n0}\tilde{\pi}_n^-, \hat{\pi}_{n0})$ is also feasible to the normalized dual problem. Furthermore, since the scaled cut associated with $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$ is tight, using a similar argument as in Part 8.3, it has the same objective value as the optimal solution $(\hat{\pi}_n^+, \hat{\pi}_n^-, \hat{\pi}_{n0})$. So, it must be optimal as well. This completes the proof.

□

The converse, however, does not hold: normalization can generate cuts that cannot be obtained from the regularized problem. In particular, Pareto-optimal cuts that are not tight at the incumbent can be attained via the normalized dual but are excluded by regularization-based approaches. We illustrate this distinction with a small example below.

Example 1. *We consider a two-stage MSIP with a scalar state variable* $x_{a(n)} \in Z_{a(n)} := [0, 3]$ *and second-stage value function*

$$Q_n(x_{a(n)}) \;:=\; \min\{\, x_n \mid 1.5\, x_n \geq x_{a(n)},\; x_n \in \{0, 1, 2, 3\}\,\}.$$

For the incumbent $\hat{x}_{a(n)} = 1$ with $\hat{\theta} = 0.1$, Figure 2 illustrates the value function together with two valid cuts, shown both in the original space and in the lifted space. The cut depicted with dashed lines is tight at the incumbent and Pareto-optimal. By contrast, the cut depicted with dotted lines is not tight at $\hat{x}_{a(n)}$ but remains Pareto-optimal. The latter cut can be obtained via normalization but not via regularization-based approaches, because its coefficients are infeasible in the regularization problem. In particular, the regularization constraint $\pi_n^+, \pi_n^- \in \Pi_n(\hat{x}_{a(n)})$ or its ϵ-approximation (33), is violated because the dotted cut is not tight.

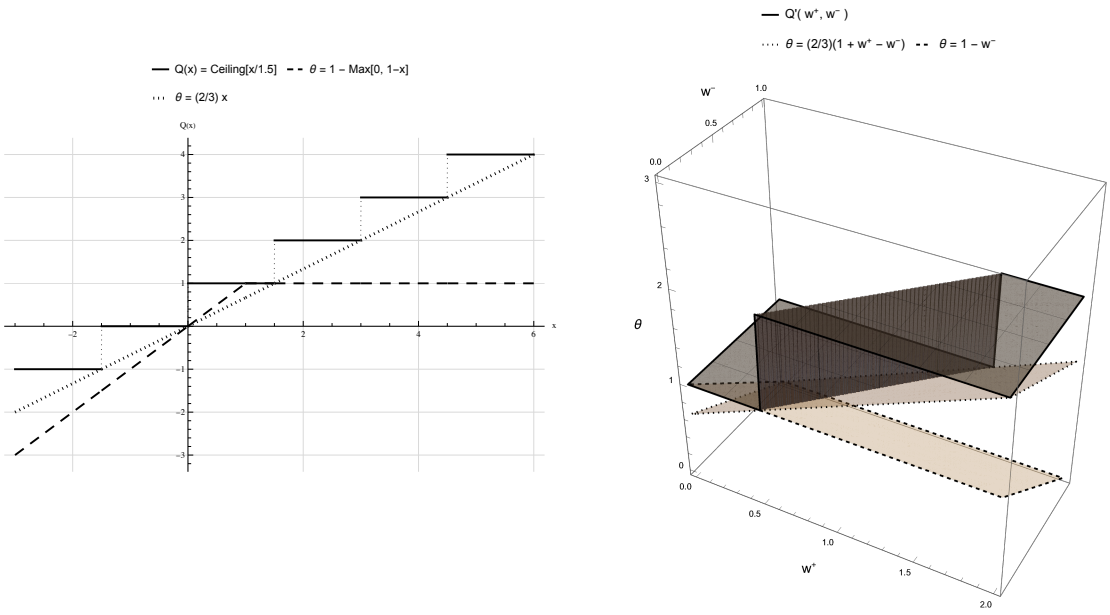

Figure 2: Left: $Q_n(\cdot)$ and the two cuts over the original domain $Z_{a(n)}$. Right: the corresponding representations in the lifted space $Z^{\text{lift}}_{\hat{x}_{a(n)}}$. The value function is shown in black; the cuts are shown with dotted and dashed boundaries.

Proposition 8 and Example 1 show that normalization provides an additional flexibility: cuts need not be both tight and Pareto-optimal simultaneously. With an appropriate choice of normalization coefficients, both properties can be enforced simultaneously. However, normalization also permits alternative choices that yield Pareto-optimal cuts that are not tight at the incumbent, while still cutting off the incumbent solution. Another advantage of normalization is that it operates on the epigraph $\text{epi}(Q_m)$, thereby allowing the incorporation of both optimality and feasibility cuts within a unified framework. In contrast, regularization-based approaches work directly with the value function Q_m and therefore rely on the assumption of relatively complete recourse. For ease of exposition, we adopt this assumption throughout the paper; however, our results extend naturally to settings in which relatively complete recourse does not hold.

Our results also extend to the classical Benders/LP setting. For classical Benders cuts, previous work (Brandenberg and Stursberg 2021, Hosseini and Turner 2025) shows that both normalization and the regularization framework of Magnanti and Wong 1981 can attain Pareto-optimal cuts. When multiple such cuts exist, however, whether normalization could achieve the same cut as regularization remained unknown. Proposition 8 resolves this question: with an appropriate choice of normalization constraints, the coefficients of any Pareto-optimal cut obtained via regularization are also optimal in the normalization dual.

4 Computational Results

In this section, we report our computational study to assess how normalization and regularization affect the strength of the generated cuts. Specifically, we compare our normalization approach with the two regularization-based methods of Deng and Xie 2024 and Yang and Yang 2025. Since neither implementation is publicly available, we reimplement both approaches.

For Deng and Xie 2024, we follow their LP-based approximation of the set $\Pi_n(\hat{x}_{a(n)})$ in (32). When this LP becomes infeasible, we add the no-good cut proposed in their paper to restore feasibility. To ensure boundedness of the LP approximation, we select the objective coefficients $(\tilde{u}_n^+, \tilde{u}_n^-)$ in (32) using their Strategy 2, which is designed to guarantee boundedness.

For Yang and Yang 2025, several algorithmic details of the level-bundle implementation are not specified. In particular, the dual model—approximated iteratively via gradient cuts—requires explicit bounds on the dual variables to prevent infeasibility or unboundedness of the approximation. We therefore impose artificial bounds chosen as sufficiently large multiples of the integer L-shaped cut coefficients. The level-bundle method also requires an initial point; we set it to the scaled L-shaped coefficients, which, in most instances, satisfy the regularization constraint (33).

We evaluate the proposed and existing methods on two-stage instances of the dynamic capacity allocation problem (DCAP) and the capacitated lot-sizing problem (CLSP), using the formulations and data-generation procedures of Deng and Xie 2024 and Füllner et al. 2024b, respectively. Our implementation and datasets are available at `https://github.com/akulbansal5/RNorm.git`. To the best of our knowledge, it is the first publicly available open-source implementation that ensures (asymptotic) convergence for general MSIPs and supports both normalization and regularization for generating strong cuts. In contrast, the SDDP package (Dowson and Kapelevich 2021) implements Lagrangian cuts that do not guarantee convergence for general MSIPs and can be weak due to dual degeneracy. The open-source implementation of Füllner et al. 2024b employs normalization to strengthen Lagrangian cuts but does not guarantee convergence for general MSIPs.

The stochastic programs and the cut-generation methods are implemented in Julia 1.9 using the JuMP package (Dunning et al. 2017). All optimization models are solved using Gurobi 12.0. Our cut-generation methods are implemented using a multi-cut strategy; we also tested a single-cut implementation, but it consistently performed worse than the multi-cut approach. The stopping criterion is triggered when one of the following conditions is met: (a) the optimality gap falls below 0.1%, (b) 5000 iterations are reached, (c) a 3600-second time limit is exceeded, or (d) no improvement larger than 10^{-9} is observed in both the lower and upper bounds for 10 consecutive iterations. Criterion (d) is included to prevent the methods from reaching the time limit when the bounds have stalled.

Both the normalized dual (17) and regularized dual (32) are solved using the level bundle method (Lemaréchal et al. 1995). To ensure a fair comparison between these approaches, we use identical parameters for the level bundle method: a convergence tolerance of 10^{-2} and a maximum of 300 iterations. The parameter ϵ in set $\Pi_n(\hat{x}_{a(n)})$ approximation (33) is also set to 10^{-2}.

The choice of the core point is made to satisfy the conditions in Proposition 4. For components k of the incumbent vector that lie on the boundary, we set $u_{nk}^+ = 10^{-3}$ and $u_{nk}^- = 0$ when $\hat{x}_{nk} = 0$. Similarly, when $\hat{x}_{nk} = B_k$, we set $u_{nk}^- = 10^{-3}$ and $u_{nk}^+ = 0$. This core point selection is identical for both the regularization and normalization methods. For the normalization method, the scalar u_{n0} is set to $\underline{Q}_n(\hat{x}_{a(n)}) - \hat{\theta}_n + \epsilon$, where $\epsilon = 10^{-6}$.

We first report the results for the DCAP instances in Table 1. The columns I, J, N, and S describe the instance characteristics: I denotes the number of resources, J the number of tasks to which resources are assigned, N the number of scenarios, and S the number of time periods. For each instance class, we randomly sample three instances and report the average results across them. All instances considered are two-stage problems; the parameter S affects the instance size but does not change the number of stages. The column "App." indicates the approach used: "Norm" for normalization, "R-LP" for LP-based regularization proposed by Deng and Xie 2024, and "Reg" for Yang and Yang 2025's regularization approach, where the dual in (32)-(33) is solved exactly using the level-bundle method. The columns "Iter" and "Time" report the number of iterations and computation time (in seconds), respectively. The columns "UB" and "LB" denote the upper bound and lower bound at termination, and "Gap (%)" reports the relative optimality gap. Column "D-Iter" denotes the average number of iterations required to solve the dual problem.

From Table 1, we observe that normalization reaches the target gap (below 0.1%) in fewer decomposition iterations ("Iter") than regularization, indicating that it produces stronger cuts and a more accurate approximation of the value function. However, the regularization approach ("Reg") typically achieves lower overall solution times. This time advantage is primarily due to the cost of solving the dual problem: the regularized dual is cheaper to solve than the normalized dual, and its associated level-bundle procedure usually requires fewer dual iterations (D-Iter). As a result, even when regularization requires more iterations (Iter) in the decomposition algorithm, it can still yield shorter overall solution times. This effect is particularly visible for $(I, J, N, S) = (3, 4, 100, 5)$ and $(4, 5, 100, 6)$, where normalization uses fewer outer iterations and fewer dual iterations, yet has a larger total solution time than regularization. We attribute this to the proximal step in the level-bundle method, which solves a quadratic program that is more difficult to solve, partly due to the additional variable π_{n0} present in the normalized dual.

Next, we report results for the CLSP instances in Table 2. The columns P and N denote the number of products and scenarios, respectively. For each (P, N) class, we consider three randomly sampled instances and report results averaged over these three instances. The columns "App.", "Iter", "D-Iter", "Time", "UB", "LB", and "Gap (%)" are defined as in the previous tables.

The results in Table 2 indicate that normalization outperforms the regularization-based approach on these instances. The method "R-LP" makes little progress within the time limit, terminating with large gaps (approximately 58%–73%) and weak lower bounds. In contrast, both "Reg" and "Norm" reduce the optimality gap by orders of magnitude and typically achieve gaps below 0.1% on most instances. Moreover, "Norm" generally requires fewer decomposition iterations and attains better gaps, suggesting that the normalized method leads to stronger cuts. For larger instances ($P \in \{10, 20\}$), both "Reg" and "Norm" become substantially more expensive and do not always reach the 0.1% threshold; nevertheless, "Norm" still tends to produce tighter final gaps than "Reg" at better or comparable run times. Unlike the DCAP results, "Norm" is also faster than "Reg" on the CLSP instances. This is mainly because, for CLSP, the normalized dual usually converges in fewer iterations ("D-Iter") than the regularized dual, and the resulting reduction in the dual solve times improves overall solution time.

For large-scale instances, solving the dual problem required to generate ReLU cuts—whether via the normalized dual (17) or a regularized dual (32)—can become a major computational bottleneck, a phenomenon widely reported in the stochastic programming literature (e.g., Zou et al. 2019, Chen and Luedtke 2022, Füllner et al. 2024b, Yang and Yang 2025). To reduce this cost, we adopt an enhancement strategy inspired by the alternating cut criterion of Angulo et al. 2016, originally proposed to enhance the integer L-shaped method (Laporte and Louveaux 1993). Rather than generating ReLU cuts at every iteration—which is expensive—we alternate them with cheaper cuts, namely Benders cuts obtained from the LP relaxation of the subproblems. In particular, we

compute a ReLU cut only when the Benders cut from the LP relaxation fails to cut off the incumbent solution. This strategy has been shown to significantly improve performance on MSIPs (see Bansal and Küçükyavuz 2024). We report results under this alternating strategy in Tables 3 and 4 for the DCAP and CLSP instances, respectively. The "Prop." column indicates the proportion of ReLU cuts among all cuts added, including Benders cuts.

Table 3 reports results for DCAP instances with the alternating cut generation approach. Comparing these results to those in Table 1, we observe significant improvements in both solution times and final optimality gaps across most instances. The "Prop." column reveals that ReLU cuts are needed in only a small fraction of iterations: for smaller instances such as $(I, J, N, S) = (2, 2, 10, 4)$ and $(2, 3, 10, 4)$, Benders cuts alone suffice (Prop. = 0.00–0.03), meaning the problem can be solved without generating expensive ReLU cuts. For larger instances, the proportion increases but remains modest, reaching at most 0.38 for $(I, J, N, S) = (4, 5, 10, 6)$ with the "Norm" approach. This translates to Benders cuts successfully cutting off the incumbent solution approximately 62%–100% of the time for the "Reg" and "Norm" approaches. The "D-Iter" values appear lower in Table 3 compared to Table 1, but this is because "D-Iter" now averages over both ReLU cut iterations (which require dual solves) and Benders cut iterations (which require 0 dual iterations). While the alternating criterion does increase the total number of decomposition iterations (e.g., for $(I, J, N, S) = (4, 5, 10, 6)$, "Reg" increases from 50 to 66 iterations and "Norm" from 28 to 53 iterations) due to the weaker Benders cuts requiring more iterations to close the gap, the computational savings from cheaper cuts per iteration result in significantly faster overall solution times. For instance, solution times for $(I, J, N, S) = (4, 5, 10, 6)$ improve from 1001 to 331 seconds for "Reg" and from 1389 to 910 seconds for "Norm", despite the increase in decomposition iterations.

The results in Table 4 show a more mixed impact of the alternating criterion on CLSP instances compared to DCAP. The "Prop." column reveals that, unlike DCAP instances, CLSP instances require ReLU cuts in a much larger fraction of iterations (typically 60%–90%), meaning Benders cuts fail to cut off the incumbent solution in most iterations. Consequently, the alternating strategy provides limited computational savings from cheaper Benders cuts, and we observe modest improvements or slight increases in solution times for smaller instances. However, for larger instances that hit the time limit, the alternating criterion combined with the normalization method yields significant improvements in final optimality gaps. For example, for $(P, N) = (20, 10)$ and $(20, 100)$, "Norm" achieves gaps of 1.91% and 3.81%, respectively, compared to gaps of 2.42% and 6.35% when the alternating criterion is not applied (as seen in Table 2). Consistent with the DCAP results, the alternating criterion increases the total number of decomposition iterations compared to the non-alternating approach (e.g., for $(P, N) = (5, 10)$, "Reg" increases from 35 to 46 iterations and "Norm" from 29 to 35 iterations), as weaker Benders cuts require more iterations to close the gap.

To summarize our computational findings, normalization consistently produces stronger cuts than regularization across both DCAP and CLSP instances, as evidenced by fewer decomposition iterations required to reach the target optimality gap. However, solving the normalized dual can be more expensive than solving the regularized dual, leading to mixed results in overall solution times depending on the problem structure. On DCAP instances, regularization often achieves faster solution times despite requiring more iterations, while on CLSP instances, normalization's advantage in both cut strength and dual convergence typically results in faster overall performance. The alternating cut criterion substantially improves performance on DCAP instances by leveraging cheaper Benders cuts, reducing solution times by up to 50% while maintaining or improving optimality gaps. On CLSP instances, where Benders cuts are less effective, the alternating criterion provides more modest benefits but still yields notable improvements in final gaps for larger instances that hit the time limit, particularly when combined with normalization.

I	J	N	S	App.	Iter	Time	UB	LB	Gap(%)	D-Iter
2	2	10	4	R-LP	17	12	1031.57	1031.33	0.021	0
				Reg	6	8	1031.41	1031.24	0.016	12
				Norm	5	16	1031.43	1030.67	0.075	14
2	2	100	4	R-LP	6	16	1094.71	1093.72	0.09	0
				Reg	6	29	1094.28	1093.32	0.088	13
				Norm	5	35	1094.37	1093.71	0.063	13
2	3	10	4	R-LP	24	13	1499.58	1496.82	0.196	0
				Reg	10	11	1497.28	1497.11	0.012	16
				Norm	7	11	1497.57	1496.95	0.041	16
2	3	100	4	R-LP	17	36	1675.32	1675.01	0.019	0
				Reg	8	53	1675.31	1674.26	0.063	15
				Norm	7	47	1675.31	1674.87	0.026	15
3	4	10	5	R-LP	162	1695	2171.00	2127.54	1.89	0
				Reg	24	104	2154.60	2154.29	0.014	30
				Norm	16	147	2154.96	2153.82	0.053	33
3	4	100	5	R-LP	102	2486	2207.11	2201.32	0.252	0
				Reg	15	527	2204.29	2203.25	0.047	28
				Norm	12	613	2204.35	2203.40	0.042	25
4	5	10	6	R-LP	214	T	2663.60	2493.28	6.341	0
				Reg	50	1001	2621.57	2619.88	0.065	46
				Norm	28	1389	2621.37	2619.74	0.062	74
4	5	100	6	R-LP	87	T	3047.66	2896.04	4.953	0
				Reg	26	2990	3011.89	2993.50	0.569	45
				Norm	21	3315	3006.23	3000.85	0.172	42

Table 1: Comparison of normalization and regularization-based approaches on DCAP instances

5 Conclusion

We studied the problem of weak (and potentially ineffective) Lagrangian cuts that arise from dual degeneracy in decomposition methods for multistage stochastic integer programs with mixed-integer state variables. Building on the ReLU-dual framework of Deng and Xie 2024 and the normalization framework of Füllner et al. 2024b, we introduced a normalized version of the ReLU dual that selects dual solutions through additional normalization constraints. This yields cut coefficients that, in practice, lead to stronger cuts and improved value-function approximations.

On the theoretical side, we established that normalized ReLU cuts can be interpreted directly in the original state space and that normalization can be used to recover strong cuts despite dual degeneracy. In particular, we introduced a notion of Pareto-optimality for nonlinear ReLU cuts in the original space and showed that normalization produces Pareto-optimal cuts under this definition. We also proved that there exists a choice of normalization coefficients that yields cuts that are tight at the current incumbent, providing guidance for selecting normalization coefficients when tightness is desired. Finally, we clarified the connection to recently proposed regularization strategies: any cut obtainable from regularization of the optimal ReLU-dual set can also be obtained via normalization (up to scaling), while normalization is strictly more flexible because it can generate Pareto-optimal cuts that are not necessarily tight at the incumbent, yet still separate the incumbent solution.

Our computational study on DCAP and CLSP instances demonstrates that the theoretical advantages of normalization translate into practical benefits: normalized cuts consistently require fewer decomposition iterations to reach the target gap, confirming their superior strength. The combina-

P	N	App.	Iter	Time	UB	LB	Gap(%)	D-Iter
3	2	R-LP	408	T	741.95	232.90	68.247	0
		Reg	17	6	583.84	583.61	0.046	14
		Norm	11	6	546.19	545.90	0.044	13
5	10	R-LP	200	T	861.80	359.72	58.248	0
		Reg	35	79	804.32	803.64	0.085	35
		Norm	29	43	804.26	803.50	0.095	21
5	50	R-LP	149	T	868.72	360.60	58.483	0
		Reg	28	248	814.40	814.01	0.048	37
		Norm	21	94	814.48	813.83	0.081	19
5	100	R-LP	130	T	866.44	360.67	58.365	0
		Reg	33	667	811.02	810.29	0.090	36
		Norm	20	164	811.03	810.37	0.081	17
10	10	R-LP	117	T	2539.38	696.84	72.521	0
		Reg	80	T	1838.53	1770.76	3.677	94
		Norm	65	T	1834.51	1801.39	1.802	57
10	50	R-LP	103	T	2552.69	697.84	72.660	0
		Reg	40	T	1845.30	1726.71	6.431	80
		Norm	35	T	1839.71	1778.45	3.342	59
10	100	R-LP	93	T	2539.51	697.98	72.514	0
		Reg	30	T	1860.03	1715.24	7.797	63
		Norm	26	T	1853.86	1770.69	4.488	57
20	10	R-LP	81	T	3900.79	1331.81	65.850	0
		Reg	56	T	3128.01	2848.67	8.930	99
		Norm	50	T	3057.22	2983.12	2.419	98
20	50	R-LP	69	T	3877.48	1331.86	65.648	0
		Reg	30	T	3150.86	2809.64	10.828	92
		Norm	24	T	3108.54	2988.41	3.861	96
20	100	R-LP	62	T	3888.87	1331.87	65.751	0
		Reg	27	T	3171.94	2793.52	11.916	92
		Norm	20	T	3160.86	2959.74	6.351	80

Table 2: Comparison of normalization and regularization-based approaches on CLSP instances

tion of normalization with the alternating cut criterion further enhances performance, particularly when Benders cuts can effectively contribute to the decomposition process. Together, these results suggest that normalization provides a flexible approach to generating strong Lagrangian cuts that can improve the efficiency of decomposition methods for multistage stochastic integer programs.

Several directions remain for future work. First, developing adaptive, instance-dependent strategies for selecting normalization coefficients—beyond the baseline rule used in our experiments—could further improve robustness and speed. Second, combining normalization with cut-selection policies and more advanced bundle-management techniques may reduce the overhead of solving the normalized dual on difficult instances.

I	J	N	S	App.	Iter	Time	UB	LB	Gap (%)	D-Iter	Prop.
2	2	10	4	R-LP	9	5	1031.4	1031.4	0	0	0.00
				Reg	9	5	1031.40	1031.40	0.000	0	0.00
				Norm	9	5	1031.40	1031.40	0.000	0	0.00
2	2	100	4	R-LP	7	7	1094.28	1094.26	0.001	0	0.00
				Reg	7	6	1094.28	1094.26	0.001	0	0.00
				Norm	7	6	1094.28	1094.26	0.000	0	0.00
2	3	10	4	R-LP	10	7	1497.27	1497.27	0	0	0.08
				Reg	10	6	1497.27	1497.27	0.000	1	0.03
				Norm	10	6	1497.27	1497.27	0.000	1	0.03
2	3	100	4	R-LP	19	19	1675.31	1675.27	0.002	0	0.42
				Reg	10	14	1675.31	1675.27	0.002	1	0.06
				Norm	10	12	1675.31	1675.27	0.000	1	0.06
3	4	10	5	R-LP	123	140	2155.27	2153.57	0.08	0	0.77
				Reg	33	40	2154.60	2153.76	0.038	6	0.26
				Norm	32	67	2154.61	2153.98	0.000	12	0.30
3	4	100	5	R-LP	59	255	2204.69	2202.38	0.105	0	0.59
				Reg	20	161	2204.33	2203.76	0.026	4	0.13
				Norm	20	176	2204.30	2203.64	0.000	4	0.13
4	5	10	6	R-LP	268	1696	2625.66	2612.68	0.49	0	0.76
				Reg	66	331	2621.65	2620.10	0.059	12	0.34
				Norm	53	910	2621.06	2619.73	0.001	39	0.38
4	5	100	6	R-LP	222	T	3006.03	2999.35	0.225	0	0.81
				Reg	46	1381	3005.08	3003.36	0.058	12	0.30
				Norm	45	3070	3006.16	3002.92	0.001	13	0.29

Table 3: Alternating criterion applied to normalization and regularization-based approaches on DCAP instances

Appendix

A Proof of Proposition 1

Recall that the subproblem $\underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$ is reformulated with copy constraints in (12), and the corresponding Lagrangian relaxation, dual, and Lagrangian cut in the lifted space are given by (13), (14), and (15), respectively. The key observation is that the Lagrangian relaxation (13) is a reformulation of the ReLU-based relaxation (6) using the linearization in (9). Specifically, by substituting the constraints from (9) into the ReLU-based relaxation, we obtain $\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) = \mathcal{L}_n^O(\pi_n^+, \pi_n^-; \hat{x}_{a(n)})$. Similarly, with the constraints $(w_n^+, w_n^-) \in Z^{lift}_{\hat{x}_{a(n)}}$ from (9), the Lagrangian cut (15) is equivalent to the ReLU cut (8). This establishes that the ReLU Lagrangian cut (8), generated at $\hat{x}_{a(n)}$ for $\text{epi}_{Z_{a(n)}}\left(\underline{Q}_n\right)$, corresponds to the Lagrangian cut (15) generated at $(\mathbf{0}, \mathbf{0})$ for the lifted epigraphical set $\text{epi}_{Z^{lift}_{\hat{x}_{a(n)}}}\left(\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)})\right)$.

For optimal dual multipliers obtained by solving the ReLU dual (7), we have $\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) = \underline{Q}_n(\hat{x}_{a(n)})$. Since $\mathcal{L}_n^R(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) = \mathcal{L}_n^O(\pi_n^+, \pi_n^-; \hat{x}_{a(n)})$ and $\underline{Q}_n(\hat{x}_{a(n)}) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$, it follows

P	N	App.	Iter	Time (s)	UB	LB	Gap (%)	D-Iter	Prop
3	2	R-LP	448	T	662.68	380.24	43.032	0	0.997
		Reg	14	7	583.87	583.68	0.034	14	0.818
		Norm	11	7	583.79	583.55	0.039	9	0.688
5	10	R-LP	238	T	861.44	429.23	50.174	0	0.995
		Reg	46	77	804.19	803.78	0.052	29	0.899
		Norm	35	57	804.24	803.59	0.080	18	0.901
5	50	R-LP	187	T	870.56	434.98	50.034	0	0.995
		Reg	34	254	814.40	813.82	0.072	29	0.877
		Norm	28	125	814.57	813.98	0.073	16	0.899
5	100	R-LP	141	T	869.15	433.32	50.136	0	0.993
		Reg	38	593	811.05	810.29	0.093	27	0.879
		Norm	26	192	811.05	810.47	0.071	14	0.872
10	10	R-LP	270	T	2046.21	971.68	52.568	0	0.986
		Reg	105	T	1840.43	1771.32	3.750	73	0.855
		Norm	71	T	1832.36	1800.79	1.723	50	0.903
10	50	R-LP	268	T	2089.39	958.68	54.101	0	0.985
		Reg	78	T	1856.18	1733.01	6.640	39	0.765
		Norm	46	T	1836.98	1779.64	3.128	43	0.825
10	100	R-LP	233	T	2113.81	959.55	54.603	0	0.980
		Reg	66	T	1863.00	1713.85	8.015	30	0.699
		Norm	40	T	1859.48	1772.68	4.670	36	0.777
20	10	R-LP	179	T	3854.72	1413.01	63.342	0	0.994
		Reg	116	T	3110.37	2857.73	8.111	47	0.581
		Norm	68	T	3045.22	2987.04	1.908	72	0.766
20	50	R-LP	77	T	3875.89	1411.27	63.585	0	0.987
		Reg	83	T	3102.65	2782.38	10.318	28	0.495
		Norm	49	T	3101.77	2990.30	3.593	49	0.689
20	100	R-LP	78	T	3886.72	1411.67	63.679	0	0.987
		Reg	75	T	3105.93	2719.35	12.448	23	0.457
		Norm	48	T	3105.17	2986.95	3.805	42	0.631

Table 4: Alternating criterion applied to normalization and regularization-based approaches on CLSP instances.

that $\mathcal{L}_n^O(\pi_n^+, \pi_n^-; \hat{x}_{a(n)}) = \underline{Q}'_n(\mathbf{0}, \mathbf{0}; \hat{x}_{a(n)})$. Therefore, the Lagrangian cut (15) is tight at $(\mathbf{0}, \mathbf{0})$ with respect to the lifted value function $\underline{Q}'_n(\cdot, \cdot; \hat{x}_{a(n)})$.

B Proof of Proposition 7

The proof is largely similar to the proof of Proposition 6. We showed in that proof that there exits a ball $B_\epsilon(\mathbf{0}, \mathbf{0}) \subseteq \text{Proj}_{w_n^+, w_n^-} \text{conv}(\mathcal{W}_{\hat{x}_{a(n)}})$. Now, choose any $(u_n^+, u_n^+) \in \text{relint}(\text{conv}(Z^{lift}_{\hat{x}_{a(n)}}))$, then we know there, exists scalar $\hat{\eta}_n$ small enough such that $\hat{\eta}_n \cdot (u_n^+, u_n^+) \in B_\epsilon(\mathbf{0}, \mathbf{0})$. Now, if we choose u_{n0} large enough then we can ensure that $\eta_n^* = \hat{\eta}_n$, and using (31), we will have $(\tilde{w}_n, \tilde{w}_n) \in B_\epsilon(\mathbf{0}, \mathbf{0})$. In particular, we let $u_{n0} = \frac{1}{\hat{\eta}_n} \left(\overline{\text{co}}(\underline{Q}'_n(u_n^+, u_n^-; \hat{x}_{a(n)})) - \hat{\theta}_n \right)$. Then, following similar reasoning as in the proof of Proposition 6, we conclude that the resulting cut is tight and Pareto-optimal.

References

Shabbir Ahmed, Filipe Goulart Cabral, and Bernardo Freitas Paulo da Costa. Stochastic Lipschitz dynamic programming. *Mathematical Programming*, 191(2):755–793, 2022.

Gustavo Angulo, Shabbir Ahmed, and Santanu S Dey. Improving the integer L-shaped method. *INFORMS Journal on Computing*, 28(3):483–499, 2016.

Akul Bansal and Simge Küçükyavuz. A computational study of cutting-plane methods for multi-stage stochastic integer programs. *arXiv preprint arXiv:2405.02533*, 2024.

Jacques F Benders. Partitioning procedures for solving mixed-variable programming problems, Numerische Matkematic 4. *SS8*, 1962.

John R Birge. Decomposition and partitioning methods for multistage stochastic linear programs. *Operations Research*, 33(5):989–1007, 1985.

Charles E Blair and Robert G Jeroslow. The value function of an integer program. *Mathematical Programming*, 23(1):237–273, 1982.

René Brandenberg and Paul Stursberg. Refined cut selection for Benders decomposition: applied to network capacity expansion problems. *Mathematical Methods of Operations Research*, 94(3): 383–412, 2021.

Rui Chen and James Luedtke. On generating Lagrangian cuts for two-stage stochastic integer programs. *INFORMS Journal on Computing*, 34(4):2332–2349, 2022.

Haoyun Deng and Weijun Xie. On the ReLU Lagrangian cuts for stochastic mixed integer programming. *arXiv preprint arXiv:2411.01229*, 2024.

Oscar Dowson and Lea Kapelevich. Sddp. jl: a Julia package for stochastic dual dynamic programming. *INFORMS Journal on Computing*, 33(1):27–33, 2021.

Iain Dunning, Joey Huchette, and Miles Lubin. Jump: A modeling language for mathematical optimization. *SIAM Review*, 59(2):295–320, 2017. doi: 10.1137/15m1020575.

Matteo Fischetti, Domenico Salvagnin, and Arrigo Zanette. A note on the selection of Benders' cuts. *Mathematical Programming*, 124:175–182, 2010.

Christian Füllner and Steffen Rebennack. Stochastic dual dynamic programming and its variants: A review. *SIAM Review*, 67(3):415–539, 2025.

Christian Füllner, X Andy Sun, and Steffen Rebennack. On Lipschitz regularization and Lagrangian cuts in multistage stochastic mixed-integer linear programming. *Available at Optimization Online*, 2024a.

Christian Füllner, X Andy Sun, and Steffen Rebennack. A new framework to generate Lagrangian cuts in multistage stochastic mixed-integer programming. *Optimization Online*, 2024b.

Dinakar Gade, Simge Küçükyavuz, and Suvrajeet Sen. Decomposition algorithms with parametric Gomory cuts for two-stage stochastic integer programs. *Mathematical Programming*, 144(1):39–64, 2014.

Mojtaba Hosseini and John Turner. Deepest cuts for Benders decomposition. *Operations Research*, 73(5):2591–2609, 2025.

Simge Küçükyavuz and Suvrajeet Sen. An introduction to two-stage stochastic mixed-integer programming. In *Leading Developments from INFORMS Communities*, pages 1–27. INFORMS, 2017.

Gilbert Laporte and François V Louveaux. The integer L-shaped method for stochastic integer programs with complete recourse. *Operations Research Letters*, 13(3):133–142, 1993.

Claude Lemaréchal, Arkadii Nemirovskii, and Yurii Nesterov. New variants of bundle methods. *Mathematical Programming*, 69:111–147, 1995.

Thomas L Magnanti and Richard T Wong. Accelerating Benders decomposition: Algorithmic enhancement and model selection criteria. *Operations Research*, 29(3):464–484, 1981.

Robert R Meyer. On the existence of optimal solutions to integer and mixed-integer programming problems. *Mathematical Programming*, 7(1):223–235, 1974.

Lewis Ntaimo. Fenchel decomposition for stochastic mixed-integer programming. *Journal of Global Optimization*, 55:141–163, 2013.

Mario VF Pereira and Leontina MVG Pinto. Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, 52:359–375, 1991.

Andrew B Philpott, Faisal Wahid, and Joseph Frédéric Bonnans. Midas: A mixed integer dynamic approximation scheme. *Mathematical Programming*, 181(1):19–50, 2020.

Yunwei Qi and Suvrajeet Sen. The ancestral Benders' cutting plane algorithm with multi-term disjunctions for mixed-integer recourse decisions in stochastic programming. *Mathematical Programming*, 161:193–235, 2017.

Ragheb Rahmaniani, Teodor Gabriel Crainic, Michel Gendreau, and Walter Rei. The Benders decomposition algorithm: A literature review. *European Journal of Operational Research*, 259(3): 801–817, 2017.

Ward Romeijnders and Niels van der Laan. Benders decomposition with scaled cuts for multistage stochastic mixed-integer programs. *Optimization Online*, 2024.

Ward Romeijndersa, Yihang Zhangb, and Suvrajeet Sen. Stochastic mixed-integer programming: A survey. *Optimization Online*, 2025.

Suvrajeet Sen and Julia L Higle. The C^3 theorem and a D^2 algorithm for large scale stochastic mixed-integer programming: Set convexification. *Mathematical Programming*, 104:1–20, 2005.

Suvrajeet Sen and Hanif D Sherali. Decomposition with branch-and-cut approaches for two-stage stochastic mixed-integer programming. *Mathematical Programming*, 106:203–223, 2006.

Paul Melvin Stursberg. *On the Mathematics of Energy System Optimization: Network Models, Decomposition, and Economic Incentives*. PhD thesis, Technische Universität München, 2019.

Niels van der Laan and Ward Romeijnders. A converging Benders' decomposition algorithm for two-stage mixed-integer recourse models. *Operations Research*, 72(5):2190–2214, 2024.

Hanbin Yang and Haoxiang Yang. Globally converging algorithm for multistage stochastic mixed-integer programs via enhanced Lagrangian cuts. *Optimization Online*, 2025.

Minjiao Zhang and Simge Küçükyavuz. Finitely convergent decomposition algorithms for two-stage stochastic pure integer programs. *SIAM Journal on Optimization*, 24(4):1933–1951, 2014.

Jikai Zou, Shabbir Ahmed, and Xu Andy Sun. Stochastic dual dynamic integer programming. *Mathematical Programming*, 175(1):461–502, 2019.