# INEXACT CUTS IN SDDP APPLIED TO MULTISTAGE STOCHASTIC NONDIFFERENTIABLE PROBLEMS

VINCENT GUIGUES*, RENATO MONTEIRO†, AND BENAR SVAITER‡

**Abstract.** In [15], an Inexact variant of Stochastic Dual Dynamic Programming (SDDP) called ISDDP was introduced which uses approximate (instead of exact with SDDP) primal dual solutions of the problems solved in the forward and backward passes of the method. That variant of SDDP was studied in [15] for linear and for differentiable nonlinear Multistage Stochastic Programs (MSPs). In this paper, we extend ISDDP to nondifferentiable MSPs. We first provide formulas for inexact cuts for value functions of convex nondifferentiable optimization problems. We then combine these cuts with SDDP to describe ISDDP for nondifferentiable MSPs and analyze the convergence of the method. More precisely, for a problem with $T$ stages, we show that for errors bounded from above by $\varepsilon$, the limit superior and limit inferior of sequences of upper and lower bounds on the optimal value of the problem are at most at distance $3\varepsilon T$ to the optimal value and that for asymptotically vanishing errors ISDDP converges to an optimal policy. Finally, we present the results of encouraging numerical experiments on a multistage nondifferentiable stochastic convex program solved using exact SDDP and the proposed inexact variant of SDDP.

**Key words.** Stochastic optimization, SDDP, Inexact cuts for value functions, Inexact SDDP.

**AMS subject classifications.** 90C15, 90C90, 90C30

## 1. Introduction.

Multistage stochastic programs (MSPs) offer a framework to model many real-life applications but are challenging to solve, see [30] for a thorough review on MSPs.

A possible approach to approximately solve such problems is to restrict the policies to be decision rules belonging to specific classes of parametric functions, see for instance [22] and references therein. In this situation, most studies have focused on classes of problems and of decision rules allowing for a reformulation of the problem (either tight or with controlled accuracy) as a tractable optimization problem, i.e., a well structured convex optimization problem. This strategy has also been used in the context of Robust Optimization where uncertain parameters are assumed to belong to convex, nonempty, compact sets (see [3] for a thorough presentation of Robust Optimization) for instance in [4].

Another approach to solve MSPs formulated using Dynamic Programming equations is to approximate the recourse functions. Two important classes of such methods are Approximate Dynamic Programming [28] and Stochastic Dual Dynamic Programming (SDDP) [25] which is a sampling-based extension of the Nested Decomposition method [7], closely related to Stochastic Decomposition [18].

Several variants of SDDP have been proposed such as CUPPS [9], ReSa [19], the Abridged Nested Decomposition [8], MIDAS [26] for monotonic Bellman functions, or risk-averse variants [17], [29], [13], [21]. For convergence analysis of the method and variants see [27],[11],[14], [2]. We also refer to [10] which explains how to take advantage of the stationarity of the underlying stochastic processes to solve MSPs

*School of Applied Mathematics, Fundação Getulio Vargas, 190 Praia de Botafogo, Rio de Janeiro, Brazil, (vincent.guigues@fgv.br). Research of this author was partially supported by CNPq grants 401371/2014-0, 204872/2018-9 and 311289/2016-9.

† School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205, USA, (renato.monteiro@isye.gatech). Research of this author was partly supported by CNPq grant 401371/2014-0.

‡ IMPA, Estrada Dona Castorina, 110, Jardim Botânico, Rio de Janeiro, Brazil. (benar@impa.br). Research of this author was partly supported by CNPq grant 401371/2014-0

with SDDP and to [15], [23] for variants which can accelerate the convergence of SDDP. In particular, in [15], an Inexact variant of SDDP called ISDDP was introduced which allows us to solve approximately the optimization subproblems of the forward and backward passes of SDDP and to increase the accuracy of the solutions of these subproblems along the iterations of the method. ISDDP can be seen as an extension to multistage and both linear and nonlinear problems of [33] where inexact cuts were combined with Benders Decomposition [6] to solve two-stage stochastic linear programs. An inexact Stochastic Dynamic Cutting Plane (another variant of SDDP solving approximately the subproblems along the iterations of the method) was also introduced in [16] to solve MSPs. For all these inexact variants, convergence can be shown for vanishing noises and numerical experiments in [33], [15] have shown that convergence can be achieved quicker with these inexact variants.

The motivation for introducing inexact cuts obtained from the approximate primal-dual solutions of the convex nonlinear subproblems generated during the course of SDDP is due to the following reasons:

(i) a convex nonlinear subproblem can take a significant amount of time or may even be impossible to be solved to high accuracy;

(ii) it is advantageous from a practical point of view to solve the initial subproblems generated by SDDP with much less accuracy than the ones generated during its late stages; in fact, the implementation presented in [15] shows that an inexact SDDP variant based on this idea outperforms exact SDDP on several instances of a portfolio problem (see also the numerical experiments in Section 6 below).

In this paper, we extend the results of [15] to the nondifferentiable case, proposing and studying Inexact SDDP for possibly nondifferentiable multistage stochastic convex programs. More precisely, the contributions of this paper are given below.

**Contributions.**

**A. Deriving formulas for inexact cuts for value functions of possibly nondifferentiable optimization problems.** An important tool in the development of inexact variants of SDDP is the computation of inexact cuts for value functions of optimization problems, i.e., affine lower bounding functions for the value function on the basis of approximate primal-dual solutions. This task can be easily achieved for value functions of linear programs, see for instance Proposition 2.1 in [15]. For nonlinear differentiable problems, the derivation of inexact cuts is given in Propositions 2.2 and 2.3 in [15] and Proposition 3.8 in [12]. However, this task is more complicated for nondifferentiable optimization problems.

We extend these results developping tools to compute inexact cuts for value functions of nondifferentiable optimization problems. Mathematically, the problem can be stated as follows. Let $\mathcal{Q} : X \to \mathbb{R}$ be the value function given by

$$\mathcal{Q}(x) = \left\{ \begin{array}{l} \min_{y \in \mathbb{R}^m} \; f(y,x) \\ y \in Y, Ay + Bx = b, g_i(y,x) \leq 0, i = 1, \ldots, p, \end{array} \right. \tag{1.1}$$

where $X \subseteq \mathbb{R}^n, Y \subseteq \mathbb{R}^m$ and where

(H0) $X$ and $Y$ are convex, closed, and nonempty sets and $f, g_i : Y \times X \to ]-\infty, +\infty]$ are proper, lower semicontinuous, convex, and possibly nondifferentiable.

Due to (H0) value function $\mathcal{Q}$ is convex and if $\bar{x} \in \mathrm{ri}(\mathrm{dom}(\mathcal{Q}))$ then $\mathcal{Q}$ is subdifferentiable at $\bar{x}$ and there exists a cut (a lower bounding affine function) for $\mathcal{Q}$ at $\bar{x}$ which coincides with $\mathcal{Q}$ at $\bar{x}$. More generally, under some assumptions, the characterization of the subdifferential of $\mathcal{Q}$ at $\bar{x} \in X$ was given in [14, Lemma 2.1] and formulas for affine lower bounding functions for $\mathcal{Q}$ were derived in [12, Proposition 3.2] on the basis of optimal primal-dual solutions to (1.1). When only approximate primal-dual solutions are available, we can only compute inexact cuts which are still lower bounding functions for the value function but which do not coincide with this function at the point $\bar{x}$ used to compute the cut. Formulas for computing inexact cuts on the basis of approximate primal-dual solutions to (1.1) were derived in [15, 12] when functions $f, g_i$ are differentiable. In this paper, we extend in Sections 2, 3 this analysis considering possibly nondifferentiable functions $f, g_i$.

**A.1).** More precisely, in Section 2 we derive inexact cuts using a reformulation of the problem that adds some variables and constraints. Such copies of (state) variables have been used to derive cuts in several publications, for instance [20]. The novelty of the cuts we derive comes from the fact that they are built on the basis of approximate primal-dual solutions and we provide the level of inexactness of the cuts, see Proposition 2.3 and Corollary 2.4. In particular, Corollary 2.4 provides cuts easier to compute than the inexact cuts from [15] and easy to interpret. Indeed, while the computation of the cuts from [15] requires solving an additional optimization problem, Corollary 2.4 provides an analytic formula for the inexact cuts with the slope being simply an approximate dual solution, the intercept being the dual problem approximate optimal value, and the level of inexactness being the difference between the approximate primal and dual optimal values. For convex problems, such copy of state variables is not needed to compute exact cuts (on the basis of exact primal-dual solutions), see [14, Lemma 2.1], but it offers a simple way to derive cuts in the inexact case.

**A.2).** Section 3 provides formulas for inexact cuts when the objective $f$ has a saddle point representation. The advantage of these cuts, compared to the cuts derived in Section 2, is that they are computed without adding additional variables and constraints.

**B. Comparison with the cuts from [15] in the differentiable case.** In the case when $f$ and $g_i$ are differentiable, we compare in Section 4 the formulas for inexact cuts from [15] and the formulas from Section 2. In particular, on the basis of characterizations of approximate $\varepsilon$-optimal primal-dual solutions, we provide upper bounds on the level of inexactness of the cuts.

**C. Inexact cuts in SDDP for nondifferentiable problems.** In Section 5, we describe ISDDP for possibly nondifferentiable MSPs combining the framework of SDDP with the inexact cuts derived in Sections 2 and 3.

**D. Convergence of Inexact SDDP for nondifferentiable problems.** In Section 5, we also study the convergence of ISDDP. A useful tool for the convergence analysis of SDDP and ISDDP is Lemma 5.2 in [11] for vanishing errors and Lemma 4.1 in [15] for bounded errors. We provide different proofs of these lemmas with slightly different assumptions (see the corresponding Lemmas 5.1 and 5.2) and derive a stronger conclusion. More precisely, one of our assumptions is stronger (the continuity of $f$ [which is satisfied when the lemmas are applied to study the convergence of ISDDP]) and two are weaker. We show the almost sure uniform convergence of the approximate Bellman functions generated by ISDDP to a continuous function which coincides with the true Bellman functions at all accumulation points of the sequences of trial points. Interestingly, as for ISDDP applied to linear programs studied in

[15], we show that for a problem with $T$ stages and errors bounded from above by $\varepsilon$, the limit superior and limit inferior of sequences of upper and lower bounds on the optimal value of the problem are at most at distance $3\varepsilon T$ to the optimal value. Finally, similarly to ISDDP for nonlinear differentiable programs developped in [15], we show the convergence of ISDDP to an optimal policy for vanishing noises.

**E. Numerical experiments.** We consider 2 instances of a nondifferentiable multistage stochastic program and solve them using both exact and inexact variants of SDDP (the one proposed in [15] and Inexact SDDP given in this paper). We also consider a solution method called `MSDDP` mixing StoDCuP from [16] and Inexact SDDP. On these experiments, the inexact variants of `MSDDP` and of SDDP developped in this paper converge quicker than (exact) SDDP.

## 2. Inexact cuts for value functions of convex optimization problems.

In the sequel, the usual scalar product in $\mathbb{R}^n$ is denoted by $\langle x, y\rangle = x^\top y$ for $x, y \in \mathbb{R}^n$. The corresponding norm is $\|x\| = \|x\|_2 = \sqrt{\langle x, x\rangle}$.

The objective of this section is to compute inexact cuts with controlled accuracy $\varepsilon$ for value functions $\mathcal{Q}$ of form (1.1) on the basis of approximate primal-dual solutions to (1.1) solved for a given $x = \bar{x}$. We will call these cuts $\varepsilon$-inexact cuts at $\bar{x}$:

Definition 2.1 ($\varepsilon$-inexact cut.). *Let $\mathcal{Q} : X \to \mathbb{R}$ be a convex function with $X$ convex, $X \subset \mathrm{ri}(\mathrm{dom}(\mathcal{Q}))$, and let $\varepsilon \geq 0$. We say that $\mathcal{C} : X \to \mathbb{R}$ is an $\varepsilon$-inexact cut for $\mathcal{Q}$ at $\bar{x} \in X$ if $\mathcal{C}$ is an affine function satisfying $\mathcal{Q}(x) \geq \mathcal{C}(x)$ for all $x \in X$ and $\mathcal{Q}(\bar{x}) - \mathcal{C}(\bar{x}) \leq \varepsilon$.*

Remark 2.1. *A 0-inexact cut for $\mathcal{Q}$ at $\bar{x}$, i.e., an $\varepsilon$-inexact cut at $\bar{x}$ with $\varepsilon = 0$ will be called an exact cut for $\mathcal{Q}$ at $\bar{x}$.*

### 2.1. Affine functions of the argument in the constraints.

We start computing inexact cuts for particular value functions $\mathcal{Q}$ where the argument of this function only appears in the constraints through affine functions of this argument. The study of this case will help us discuss the general case of a value function of form (1.1) considered in the next Section 2.2.

More precisely, we consider value functions $\mathcal{Q}$ of form:

$$\mathcal{Q}(x) = \left\{ \begin{array}{l} \displaystyle\min_{y \in \mathbb{R}^m} \; f(y) \\ g(y) \leq Cx, \\ Ay + Bx = b, \\ y \in Y, \end{array} \right. \tag{2.2}$$

along with the corresponding dual problem given by

$$\left\{ \begin{array}{l} \displaystyle\max_{\lambda, \mu} \; \theta_x(\lambda, \mu) \\ \mu \geq 0, \lambda, \end{array} \right. \tag{2.3}$$

where dual function $\theta_x(\lambda, \mu)$ is given by

$$\theta_x(\lambda, \mu) = \min\{L_x(y, \lambda, \mu) : y \in Y\} \tag{2.4}$$

for the Lagrangian

$$L_x(y, \lambda, \mu) = f(y) + \langle \lambda, Ay + Bx - b\rangle + \langle \mu, g(y) - Cx\rangle.$$

Proposition 2.2 provides a formula for computing inexact cuts for value function $\mathcal{Q}$ given by (2.2):

PROPOSITION 2.2. *Assume that $f : \mathbb{R}^m \rightarrow ]-\infty, +\infty]$ and component functions $g_i : \mathbb{R}^m \rightarrow ]-\infty, +\infty]$, $i = 1, \ldots, p$, of $g$ are proper, convex, and lower semicontinuous. Assume that $\hat{y}$ is an $\varepsilon_P$-optimal feasible solution of problem (2.2) for $x = \bar{x}$ and that $(\hat{\lambda}, \hat{\mu})$ is an $\varepsilon_D$-optimal feasible solution of the corresponding dual problem (2.3) for $x = \bar{x}$. Assume that $f$ is finite on $\{y \in Y : Ay + B\bar{x} = b, g(y) \leq C\bar{x}\}$ and that Slater constraint qualification holds for (2.2) written for $x = \bar{x}$, i.e., there is $y_{\bar{x}} \in ri(Y)$, such that $Ay_{\bar{x}} + B\bar{x} = b$, $g(y_{\bar{x}}) < C\bar{x}$. Then*

$$\mathcal{C}(x) = f(\hat{y}) - (\varepsilon_P + \varepsilon_D) + \langle B^\top \hat{\lambda} - C^\top \hat{\mu}, x - \bar{x} \rangle$$

*is an $(\varepsilon_P + \varepsilon_D)$-inexact cut for $\mathcal{Q}$ at $\bar{x}$.*

*Proof.* By definition of $\hat{y}$, we get

$$f(\hat{y}) \leq \mathcal{Q}(\bar{x}) + \varepsilon_P. \tag{2.5}$$

The assumptions of the Convex Duality theorem are satisfied for problem (2.2) and its dual (2.3), both written for $x = \bar{x}$. Therefore the optimal value of dual problem (2.3) written for $x = \bar{x}$ is the optimal value $\mathcal{Q}(\bar{x})$ of the corresponding primal problem. Using the definition of $\hat{\lambda}, \hat{\mu}$, it follows that

$$\theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}) \geq \mathcal{Q}(\bar{x}) - \varepsilon_D. \tag{2.6}$$

Next,

$$\begin{array}{lcl}
\mathcal{Q}(x) & \geq & \theta_x(\hat{\lambda}, \hat{\mu}) \text{ by weak duality and feasibility of } \hat{\mu}, \hat{\lambda}, \\
& = & \min\{L_x(y, \hat{\lambda}, \hat{\mu}) : y \in Y\}, \\
& = & \langle \hat{\lambda}, B(x - \bar{x}) \rangle + \langle \hat{\mu}, -C(x - \bar{x}) \rangle + \min\{L_{\bar{x}}(y, \hat{\lambda}, \hat{\mu}) : y \in Y\}, \\
& = & \langle B^\top \hat{\lambda} - C^\top \hat{\mu}, x - \bar{x} \rangle + \theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}), \\
& \overset{(2.6)}{\geq} & \langle B^\top \hat{\lambda} - C^\top \hat{\mu}, x - \bar{x} \rangle + \mathcal{Q}(\bar{x}) - \varepsilon_D, \\
& \overset{(2.5)}{\geq} & \mathcal{C}(x) := \langle B^\top \hat{\lambda} - C^\top \hat{\mu}, x - \bar{x} \rangle + f(\hat{y}) - \varepsilon_P - \varepsilon_D.
\end{array}$$

Moreover, since $f(\hat{y}) \geq \mathcal{Q}(\bar{x})$, we get

$$\mathcal{Q}(\bar{x}) - \mathcal{C}(\bar{x}) = \varepsilon_P + \varepsilon_D + \mathcal{Q}(\bar{x}) - f(\hat{y}) \leq \varepsilon_P + \varepsilon_D,$$

and we have shown that $\mathcal{C}$ is an $(\varepsilon_P + \varepsilon_D)$-inexact cut for $\mathcal{Q}$ at $\bar{x}$. □

REMARK 2.2. *The proof of Proposition 2.2 also shows that if $\theta_{\bar{x}}(\hat{\lambda}, \hat{\mu})$ can be computed exactly (i.e., if optimization problem (2.4) written for $x = \bar{x}, \lambda = \hat{\lambda}, \mu = \hat{\mu}$ is solved to optimality) then $\mathcal{C}(x) = \theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}) + \langle B^\top \hat{\lambda} - C^\top \hat{\mu}, x - \bar{x} \rangle$ is an $\varepsilon_D$-inexact cut for $\mathcal{Q}$ at $\bar{x}$.*

**2.2. General value functions.** We now consider general value functions of form

$$\mathcal{Q}(x) = \left\{ \begin{array}{l} \displaystyle\min_{y \in \mathbb{R}^m} f(y, x) \\ g(y, x) \leq 0, \\ Ay + Bx = b, \\ y \in Y. \end{array} \right. \tag{2.7}$$

Analyzing the proof of Proposition 2.2 dedicated to the special case of value functions of form (2.2), we observe that the linearity in $x$ of Lagrangian function $L$ was crucial to derive our formula for inexact cuts. The Lagrangian obtained dualizing coupling constraints in problem (2.7) does not satisfy this property anymore. However, we can reformulate equivalently the problem in such a way that the Lagrangian of the reformulated problem satisfies this property. This reformulation is obtained adding variable $z \in \mathbb{R}^n$ together with the constraint $z = x$. We obtain the equivalent representation of problem (2.7) under the form

$$\mathcal{Q}(x) = \left\{ \begin{array}{l} \displaystyle\min_{y \in \mathbb{R}^m, z \in \mathbb{R}^n} \; f(y, z) \\ g(y, z) \leq 0, \\ Ay + Bz = b, \\ y \in Y, \\ z = x. \end{array} \right. \tag{2.8}$$

The use of the copy $z = x$ of state variables to derive cuts in the context of SDDP has been used in several publications, for instance [20, 31]. This copy of state variables adds variables and constraints and is not necessary for convex problems, even for general value functions (1.1) having nonlinear coupling constraints, see Lemma 2.1 in [14] for an analytic formula for the corresponding exact cuts. However, the use of copy of state variables offers a simple way to derive inexact cuts in the convex case, see the corresponding Proposition 2.3 and Corollary 2.4 as well as the more complicated computations of Section 3 that do not use these copies of variables but use a saddle point representation of the objective. Denoting by $S$ the set

$$S = \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : g(y, z) \leq 0, Ay + Bz = b, y \in Y\}, \tag{2.9}$$

and dualizing the coupling constraint $z = x$ in problem (2.8), we obtain the dual problem given by

$$\left\{ \begin{array}{l} \displaystyle\max_{\lambda} \; \theta_x(\lambda) \\ \lambda \in \mathbb{R}^n, \end{array} \right. \tag{2.10}$$

where dual function $\theta_x(\lambda)$ is given by

$$\theta_x(\lambda) = \min\{L_x(y, z, \lambda) : (y, z) \in S\} \tag{2.11}$$

now for the Lagrangian

$$L_x(y, z, \lambda) = f(y, z) + \langle \lambda, x - z \rangle,$$

which, as in the special case considered in the previous section, is a linear function of $x$. Therefore, for every $x, \bar{x} \in X$, for every $(y, z) \in S$, and $\lambda$, we have

$$L_x(y, z, \lambda) = \langle \lambda, x - \bar{x} \rangle + L_{\bar{x}}(y, z, \lambda)$$

and the optimal value $\theta_x(\lambda)$ of problem (2.11) is the sum of $\langle \lambda, x - \bar{x} \rangle$ and of $\theta_{\bar{x}}(\lambda)$. Observing that from Weak Duality $\theta_x(\lambda)$ is a lower bound on $\mathcal{Q}(x)$, this sum is an affine function of $x$ which is a lower bounding function for $\mathcal{Q}$. It can be bounded from below in terms of a computable affine function (which therefore is an inexact cut for $\mathcal{Q}$ at $\bar{x}$) using an approximate primal-dual solution if problem (2.7) and its dual (2.10) written for $x = \bar{x}$ satisfy the Slater assumption.

The details of these computations are given in the proof of Proposition 2.3 below which provides formulas for inexact cuts for value function (2.7). The proof of the proposition is given for completeness but, due to our previous observations, it is similar to the proof of Proposition 2.2.

Proposition 2.3. *Let Assumption (H0) hold. Assume that $\hat{y}$ is an $\varepsilon_P$-optimal feasible solution of problem (2.7) for $x = \bar{x}$ and that $\hat{\lambda}$ is an $\varepsilon_D$-optimal feasible solution of dual problem (2.10) written for $x = \bar{x}$. Assume that $f(\cdot, \bar{x})$ is finite on $\{y \in Y : Ay + b\bar{x} = b, g(y, \bar{x}) \leq 0\}$ and that the following Slater constraint qualification holds for (2.7) written for $x = \bar{x}$:*

$$\exists y_{\bar{x}} \text{ such that } (y_{\bar{x}}, \bar{x}) \in ri(S) \tag{2.12}$$

*where $S$ is given by (2.9). Then*

$$\mathcal{C}(x) = f(\hat{y}, \bar{x}) - (\varepsilon_P + \varepsilon_D) + \langle \hat{\lambda}, x - \bar{x} \rangle$$

*is an $(\varepsilon_P + \varepsilon_D)$-inexact cut for $\mathcal{Q}$ at $\bar{x}$.*

*Proof.* By definition of $\hat{y}$, we get

$$f(\hat{y}, \bar{x}) \leq \mathcal{Q}(\bar{x}) + \varepsilon_P.$$

The assumptions of the Convex Duality theorem for dual problem (2.10) and primal problem (2.7) written for $x = \bar{x}$ are satisfied and therefore the optimal value of dual problem (2.10) written for $x = \bar{x}$ is the optimal value $\mathcal{Q}(\bar{x})$ of the corresponding primal problem. Therefore, using the definition of $\hat{\lambda}$, we get

$$\theta_{\bar{x}}(\hat{\lambda}) \geq \mathcal{Q}(\bar{x}) - \varepsilon_D.$$

Next,

$$\begin{array}{lll}
\mathcal{Q}(x) & \geq & \theta_x(\hat{\lambda}) \text{ by weak duality and feasibility of } \hat{\lambda}, \\
& = & \min\{L_x(y, z, \hat{\lambda}) : (y, z) \in S\}, \\
& = & \langle \hat{\lambda}, x - \bar{x} \rangle + \min\{L_{\bar{x}}(y, z, \hat{\lambda}) : (y, z) \in S\}, \\
& = & \langle \hat{\lambda}, x - \bar{x} \rangle + \theta_{\bar{x}}(\hat{\lambda}), \\
& \geq & \langle \hat{\lambda}, x - \bar{x} \rangle + \mathcal{Q}(\bar{x}) - \varepsilon_D, \\
& \geq & \mathcal{C}(x) := \langle \hat{\lambda}, x - \bar{x} \rangle + f(\hat{y}, \bar{x}) - \varepsilon_P - \varepsilon_D.
\end{array}$$

Moreover, since $f(\hat{y}, \bar{x}) \geq \mathcal{Q}(\bar{x})$, we get

$$\mathcal{Q}(\bar{x}) - \mathcal{C}(\bar{x}) = \varepsilon_P + \varepsilon_D + \mathcal{Q}(\bar{x}) - f(\hat{y}, \bar{x}) \leq \varepsilon_P + \varepsilon_D,$$

which achieves the proof. ∎

As before, observe that if $\theta_{\bar{x}}(\hat{\lambda})$ is available, i.e., if optimization problem (2.11) written for $x = \bar{x}$ and $\lambda = \hat{\lambda}$ is solved to optimality then $\langle \hat{\lambda}, x - \bar{x} \rangle + \theta_{\bar{x}}(\hat{\lambda})$ is an $\varepsilon_D$-inexact cut for $\mathcal{Q}$ at $\bar{x}$.

We also have the following corollary of Proposition 2.3 that will be used in the numerical simulations of Section 6, offering an inexact cut easy to implement as long as we have access to approximate primal-dual solutions:

Corollary 2.4. *Under the assumptions of Proposition 2.3, let $\hat{y}$ be any approximate optimal and feasible solution of primal problem (2.7) for $x = \bar{x}$ and let $\hat{\lambda}$ be any approximate optimal feasible solution of dual problem (2.10) written for $x = \bar{x}$. Then*

$$\mathcal{C}(x) = \theta_{\bar{x}}(\hat{\lambda}) + \langle \hat{\lambda}, x - \bar{x} \rangle$$

*is an $(f(\hat{y}, \bar{x}) - \theta_{\bar{x}}(\hat{\lambda}))$-inexact cut for $\mathcal{Q}$ at $\bar{x}$. When $\hat{y}$ and $\hat{\lambda}$ are optimal solutions then we get, as expected, an exact cut since $f(\hat{y}, \bar{x}) = \theta_{\bar{x}}(\hat{\lambda}) = \mathcal{Q}(\bar{x})$.*

*Proof.* It suffices to observe that $\hat{y}$ is an $\varepsilon_P$ optimal primal solution with $\varepsilon_P = f(\hat{y}, \bar{x}) - \mathcal{Q}(\bar{x})$, that $\hat{\lambda}$ is an $\varepsilon_D$ optimal dual solution with $\varepsilon_D = \mathcal{Q}(\bar{x}) - \theta_{\bar{x}}(\hat{\lambda})$ and to apply Proposition 2.3. □

It is also worth mentioning that if we have access to an optimal primal-dual solution to (2.7) then we can obtain an exact cut for $\mathcal{Q}$ at $\bar{x}$ directly solving (2.7) and its dual, without adding constraint $z = x$. More precisely, a characterization of the subdifferentiable of $\mathcal{Q}$ and formulas for exact cuts for $\mathcal{Q}$ given by (2.7) can be found in Lemma 2.1 in [14] and Proposition 3.2 in [12].

## 3. Inexact cuts for value functions with saddle point representation of the objective.

The inexact cuts proposed in this section are based on the observation that many convex functions have saddle point representations, see for instance [24] and Section 5.6.1.1 in [5]. More precisely, we assume that the objective function $f$ has a saddle point representation: if $p = (y, x)$, function $f$ is given by

$$f(p) = p^T a + \max_{w \in \mathcal{W}} \left[ p^T C_0 w - \phi_0(w) \right] \tag{3.13}$$

for some known convex, proper, lower semicontinuous function $\phi_0$, some known convex, compact, nonempty set $\mathcal{W}$, vector $a$, and matrix $C_0$. In this situation, we will derive inexact cuts for $\mathcal{Q}$ without additional variables $z \in \mathbb{R}^n$ and constraints $z = x$ introduced in the previous section.

"Well structured" convex functions have saddle point representations, see for instance [24] and Section 5.6.1.1 in [5] for details.

EXAMPLE 3.1. *Function $f(p) = f(y, x) = \|y - x\|_1$ has the saddle point representation $f(p) = f(y, x) = \|y - x\|_1 = \max_{\|w\|_\infty \leq 1} [w^T y - w^T x]$ which is of form (3.13) with $\mathcal{W} = \{w : \|w\|_\infty \leq 1\}$, $C_0 = [I; -I]$, and $\phi_0$ the null function.*

We start considering value functions of form

$$\mathcal{Q}(x) = \left\{ \begin{array}{l} \min\limits_{y \in \mathbb{R}^m} f(y, x) \\ y \in Y \end{array} \right. \tag{3.14}$$

with $Y$ compact, convex, and nonempty. Let $a = [a_2; a_1]$ and let us write matrix $C_0 = [A_0; B_0]$ where $A_0$ contains the first $m$ rows and $B_0$ the last $n$ rows of $C_0$. Representation (3.13) can then be written

$$f(y, x) = x^T a_1 + y^T a_2 + \max_{w \in \mathcal{W}} y^T A_0 w + x^T B_0 w - \phi_0(w) \tag{3.15}$$

and problem (3.14) becomes the saddle point problem

$$\mathcal{Q}(x) = \min_{y \in Y} \max_{w \in \mathcal{W}} \; x^T a_1 + y^T a_2 + y^T A_0 w + x^T B_0 w - \phi_0(w). \tag{3.16}$$

Since $Y$ and $\mathcal{W}$ are convex, compact and nonempty, this saddle point problem can be equivalently written as the convex problem

$$\mathcal{Q}(x) = x^T a_1 + \left\{ \begin{array}{l} \max\limits_{w} \; \theta_x(w) \\ w \in \mathcal{W} \end{array} \right. \tag{3.17}$$

where concave function $\theta_x$ is given by

$$\theta_x(w) = \left\{ \begin{array}{l} \min\limits_{y} \; L_x(y, w) \\ y \in Y, \end{array} \right. \tag{3.18}$$

where

$$L_x(y,w) = y^T(a_2 + A_0 w) + x^T B_0 w - \phi_0(w). \tag{3.19}$$

Once again, the linearity in $x$ of this new Lagrangian function $L_x(y,w)$ will allow us to derive inexact cuts. However, contrary to the previous section, this linearity was achieved using a saddle point representation of $f$. The following proposition provides inexact cuts for $\mathcal{Q}$ given by (3.14) with $f$ of the form (3.15).

Proposition 3.2. *Consider problem (3.14) with $f$ having a saddle point representation of form (3.15). Assume that $Y$ and $\mathcal{W}$ are compact, convex, and nonempty. Let $\hat{w} \in \mathcal{W}$ be an $\varepsilon$-optimal solution of problem (3.17) written with $x = \bar{x}$ and let $\hat{y} \in Y$ be a $\tau$-optimal solution of problem (3.18) written with $x = \bar{x}, w = \hat{w}$. Then the affine function*

$$\mathcal{C}(x) := x^\top \Big( a_1 + B_0 \hat{w} \Big) + \hat{y}^\top \Big( a_2 + A_0 \hat{w} \Big) - \phi_0(\hat{w}) - \tau \tag{3.20}$$

*is a $(\varepsilon + \tau)$-inexact cut for $\mathcal{Q}$ at $\bar{x}$.*

*Proof.* Let $(\bar{y}, \bar{w})$ be an optimal solution of saddle point problem (3.16) with $x = \bar{x}$. By definition of $\hat{w}$ and $\hat{y}$, we have

$$\theta_{\bar{x}}(\bar{w}) - \varepsilon \le \theta_{\bar{x}}(\hat{w}) \text{ and } \theta_{\bar{x}}(\hat{w}) + \tau \ge L_{\bar{x}}(\hat{y}, \hat{w}) \ge \theta_{\bar{x}}(\hat{w}). \tag{3.21}$$

By linearity of $L_{\cdot}(y,w)$ we get for every $y \in Y, w \in \mathcal{W}$, that

$$L_x(y,w) = L_{\bar{x}}(y,w) + (x - \bar{x})^T B_0 w. \tag{3.22}$$

Next, using representation (3.17) of $\mathcal{Q}$ and the fact that $\hat{w} \in \mathcal{W}$ we have

$$\begin{array}{rcl} \mathcal{Q}(x) & \ge & x^T a_1 + \theta_x(\hat{w}) \\ & = & x^T a_1 + \left\{ \begin{array}{l} \min \ L_x(y, \hat{w}) \\ y \in Y, \end{array} \right. \\ & \overset{(3.22)}{=} & x^T a_1 + (x-\bar{x})^T B_0 \hat{w} + \left\{ \begin{array}{l} \min \ L_{\bar{x}}(y, \hat{w}) \\ y \in Y, \end{array} \right. \\ & = & x^T a_1 + (x-\bar{x})^T B_0 \hat{w} + \theta_{\bar{x}}(\hat{w}) \\ & \overset{(3.21)}{\ge} & x^T a_1 + (x-\bar{x})^T B_0 \hat{w} + L_{\bar{x}}(\hat{y}, \hat{w}) - \tau \\ & \overset{(3.20)}{=} & \mathcal{C}(x). \end{array}$$

Moreover,

$$0 \le \mathcal{Q}(\bar{x}) - \mathcal{C}(\bar{x}) = \tau + \theta_{\bar{x}}(\bar{w}) - L_{\bar{x}}(\hat{y}, \hat{w}) \overset{(3.18)}{\le} \tau + \theta_{\bar{x}}(\bar{w}) - \theta_{\bar{x}}(\hat{w}) \le \tau + \varepsilon,$$

which achieves the proof of the proposition. ∎

Now consider value function $\mathcal{Q}$ given by

$$\mathcal{Q}(x) = \left\{ \begin{array}{l} \min f(y,x) \\ y \in Y, \ Ay + Bx = b, \end{array} \right. \tag{3.23}$$

with $Y$ convex, nonempty, and compact. If $f$ has a saddle point representation of form (3.15) with $\mathcal{W}$ convex, nonempty, and compact, value function (3.23) can be written

$$\mathcal{Q}(x) = x^T a_1 + \left\{ \begin{array}{l} \max \theta_x(w) \\ w \in \mathcal{W} \end{array} \right. \tag{3.24}$$

where

$$\theta_x(w) = \left\{ \begin{array}{l} \min\ y^T(a_2 + A_0 w) + x^T B_0 w - \phi_0(w) \\ y \in Y, Ay + Bx = b. \end{array} \right. \tag{3.25}$$

For problem (3.25) define the Lagrangian

$$\begin{array}{rcl} \mathcal{L}_{x,w}(y,\lambda) & = & y^T(a_2 + A_0 w) + x^T B_0 w - \phi_0(w) + \lambda^T(Ay + Bx - b) \\ & = & L_x(y,w) + \lambda^T(Ay + Bx - b), \end{array} \tag{3.26}$$

where $L_x(y,w)$ is given by (3.19). Let us fix $\bar{x} \in \mathbb{R}^n$ and assume that there is $y_0 \in \mathrm{ri}(Y)$ such that $Ay_0 + B\bar{x} = b$. Then by the Convex Duality theorem, we can express $\theta_{\bar{x}}(w)$ as the optimal value of the dual of (3.25):

$$\theta_{\bar{x}}(w) = \max_{\lambda}\ h_{\bar{x},w}(\lambda) \tag{3.27}$$

for the dual function

$$h_{\bar{x},w}(\lambda) = \left\{ \begin{array}{l} \min\ \mathcal{L}_{\bar{x},w}(y,\lambda) \\ y \in Y. \end{array} \right. \tag{3.28}$$

Proposition 3.3. *Consider problem (3.23) with $f$ having a saddle point representation of form (3.15). Assume that sets $Y$ and $\mathcal{W}$ are nonempty, convex, and compact. Let us fix $\bar{x} \in \mathbb{R}^n$ and assume that there is $y_0 \in ri(Y)$ such that $Ay_0 + B\bar{x} = b$. Let $(\bar{y}, \bar{w})$ be an optimal solution of saddle point problem (3.24) with $x = \bar{x}$ and let $\hat{w} \in \mathcal{W}$ be an $\varepsilon$-optimal solution of problem (3.24) written with $x = \bar{x}$:*

$$\theta_{\bar{x}}(\hat{w}) \geq \theta_{\bar{x}}(\bar{w}) - \varepsilon, \tag{3.29}$$

*and let $\hat{\lambda} \in Y$ be a $\delta$-optimal solution of problem*

$$\theta_{\bar{x}}(\hat{w}) = \max_{\lambda}\ h_{\bar{x},\hat{w}}(\lambda)$$

*i.e.,*

$$h_{\bar{x},\hat{w}}(\hat{\lambda}) \geq \theta_{\bar{x}}(\hat{w}) - \delta. \tag{3.30}$$

*Let $\hat{y}$ be a $\tau$-optimal feasible solution of*

$$\theta_{\bar{x}}(\hat{w}) = \left\{ \begin{array}{l} \min\ y^T(a_2 + A_0 \hat{w}) + \bar{x}^T B_0 \hat{w} - \phi_0(\hat{w}) \\ y \in Y, Ay + B\bar{x} = b, \end{array} \right.$$

*i.e.,*

$$\hat{y} \in Y,\ A\hat{y} + B\bar{x} = b,\ L_{\bar{x}}(\hat{y}, \hat{w}) \leq \theta_{\bar{x}}(\hat{w}) + \tau. \tag{3.31}$$

*Then the affine function*

$$\mathcal{C}(x) = x^T\Big(a_1 + B_0\hat{w} + B^T\hat{\lambda}\Big) + \hat{y}^T\Big(a_2 + A_0\hat{w}\Big) - \bar{x}^T B^T \hat{\lambda} - \phi_0(\hat{w}) - \tau - \delta \tag{3.32}$$

*is a $(\varepsilon + \tau + \delta)$-inexact cut for $\mathcal{Q}$ at $\bar{x}$.*

*Proof.* By linearity of $\mathcal{L}_{\cdot,w}(y,\lambda)$ we get for every $y \in Y, w \in \mathcal{W}$, that

$$\mathcal{L}_{x,w}(y,\lambda) = \mathcal{L}_{\bar{x},w}(y,\lambda) + (x-\bar{x})^T(B_0 w + B^T\lambda). \tag{3.33}$$

Next, using representation (3.24) of $\mathcal{Q}$ and the fact that $\hat{w} \in \mathcal{W}$ we have

$$\begin{array}{lll}
\mathcal{Q}(x) & \geq & x^T a_1 + \theta_x(\hat{w}) \\
& \geq & x^T a_1 + h_{x,\hat{w}}(\hat{\lambda}), \\
& \overset{(3.28)}{=} & x^T a_1 + \left\{ \begin{array}{l} \min\ \mathcal{L}_{x,\hat{w}}(y,\hat{\lambda}) \\ y \in Y, \end{array} \right. \\
& \overset{(3.33)}{=} & x^T a_1 + (x-\bar{x})^T(B_0\hat{w} + B^T\hat{\lambda}) + \left\{ \begin{array}{l} \min\ \mathcal{L}_{\bar{x},\hat{w}}(y,\hat{\lambda}) \\ y \in Y, \end{array} \right. \\
& \overset{(3.28)}{=} & x^T a_1 + (x-\bar{x})^T(B_0\hat{w} + B^T\hat{\lambda}) + h_{\bar{x},\hat{w}}(\hat{\lambda}) \\
& \overset{(3.30)}{\geq} & x^T a_1 + (x-\bar{x})^T(B_0\hat{w} + B^T\hat{\lambda}) + \theta_{\bar{x}}(\hat{w}) - \delta \\
& \overset{(3.31)}{\geq} & x^T a_1 + (x-\bar{x})^T(B_0\hat{w} + B^T\hat{\lambda}) + L_{\bar{x}}(\hat{y},\hat{w}) - \tau - \delta \\
& \overset{(3.32)}{=} & \mathcal{C}(x).
\end{array}$$

Moreover, if $\bar{w}$ is an optimal solution of (3.24) written for $x = \bar{x}$, i.e., $\mathcal{Q}(\bar{x}) = \bar{x}^T a_1 + \theta_{\bar{x}}(\bar{w})$ we obtain

$$0 \leq \mathcal{Q}(\bar{x}) - \mathcal{C}(\bar{x}) = \tau + \delta + \theta_{\bar{x}}(\bar{w}) - L_{\bar{x}}(\hat{y},\hat{w}) \overset{(3.25)}{\leq} \tau + \delta + \theta_{\bar{x}}(\bar{w}) - \theta_{\bar{x}}(\hat{w}) \overset{(3.29)}{\leq} \tau + \delta + \varepsilon,$$

which achieves the proof of the proposition. □

## 4. Particular case of differentiable problems and comparison with the inexact cuts from [15].

The following proposition, taken from [15], provides an inexact cut for $\mathcal{Q}$ given by (2.7) when functions $f, g_i$ are differentiable.

PROPOSITION 4.1. *Consider value function $\mathcal{Q}$ given by (2.7). Let Assumption (H0) hold, take $\bar{x} \in X$, and assume that*

$$\text{there exists } y_{\bar{x}} \in ri(Y) \text{ such that } Ay_{\bar{x}} + B\bar{x} = b \text{ with } g(y_{\bar{x}}, \bar{x}) < 0. \tag{4.34}$$

*Assume that $f$ and $g$ are differentiable on $Y \times X$. Let $\varepsilon \geq 0$, let $\hat{y}$ be an $\epsilon$-optimal feasible primal solution for problem (2.7) written for $x = \bar{x}$ and let $(\hat{\lambda}, \hat{\mu})$ be an $\epsilon$-optimal feasible solution of the corresponding dual problem given by*

$$\max_{\mu \geq 0, \lambda} \theta_x(\lambda, \mu)$$

*where the dual function $\theta_x(\lambda,\mu)$ is given by*

$$\theta_x(\lambda,\mu) = \min_{y \in Y} L_x(y,\lambda,\mu) \tag{4.35}$$

*for the Lagrangian*

$$L_x(y,\lambda,\mu) = f(y,x) + \langle \lambda, Bx + Ay - b \rangle + \langle \mu, g(y,x) \rangle.$$

*Assume that $f(\cdot,\bar{x})$ is finite on*

$$S(\bar{x}) = \{y \in Y : Ay + B\bar{x} = b, g(y,\bar{x}) \leq 0\} \tag{4.36}$$

*and that* $\eta(\varepsilon) = \ell(\hat{y}, \bar{x}, \hat{\lambda}, \hat{\mu})$ *is finite where*

$$\ell(\hat{y}, \bar{x}, \hat{\lambda}, \hat{\mu}) = \max\{\langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y \rangle : y \in Y\}.$$

*Then the affine function*

$$\mathcal{C}(x) := L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) - \eta(\varepsilon) + \langle \nabla_x L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), x - \bar{x} \rangle \tag{4.37}$$

*is an* $(\varepsilon + \ell(\hat{y}, \bar{x}, \hat{\lambda}, \hat{\mu}))$*-inexact cut for* $\mathcal{Q}$ *at* $\bar{x}$*.*

We want to compare the inexact cuts given by Propositions 2.3 and 4.1 obtained taking $\varepsilon_D = \varepsilon_P = \varepsilon$ in Proposition 2.3. For the cut given by Proposition 4.1 to be valid, we assume that the assumptions of this proposition are satisfied. In particular, (4.34) holds. Let us show that if in addition $Y \times X \subset \mathrm{dom}(g_i)$ for all $i = 1, \ldots, p$, this implies that (2.12) holds which will imply that the assumptions of Proposition 2.3 are also satisfied and the inexact cut given by that proposition is valid. Indeed, write set $S$ given by (2.9) as $S = S_1 \cap S_2 \cap (Y \times \mathbb{R}^n)$ where $S_1 = \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : g(y, z) \leq 0\}$ and $S_2 = \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : Ay + Bz = b\}$. We have that $\mathrm{ri}(S_2) = S_2$ and

$$\mathrm{ri}(\{g_i \leq 0\}) = \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : (y, z) \in \mathrm{ri}(\mathrm{dom}(g_i)), g_i(y, z) < 0, i = 1, \ldots, p\}. \tag{4.38}$$

Since $Y \times \{\bar{x}\} \subset \mathrm{dom}(g_i), i = 1, \ldots, p$, we have $\mathrm{ri}(Y) \times \{\bar{x}\} \subset \mathrm{ri}(\mathrm{dom}(g_i)), i = 1, \ldots, p$, implying that set $\cap_{i=1}^p \mathrm{ri}(\{g_i \leq 0\})$ is nonempty since it contains the nonempty set $\mathrm{ri}(Y) \times \{\bar{x}\}$ (this set contains $(y_{\bar{x}}, \bar{x})$). Therefore $\mathrm{ri}(S_1) = \cap_{i=1}^p \mathrm{ri}(\{g_i \leq 0\}) = \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : (y, z) \in \mathrm{ri}(\mathrm{dom}(g_i)), g_i(y, z) < 0, i = 1, \ldots, p\}$. It follows that convex sets $S_1, S_2$, and $Y \times \mathbb{R}^n$ are convex and satisfy $\mathrm{ri}(S_1) \cap \mathrm{ri}(S_2) \cap (\mathrm{ri}(Y) \times \mathbb{R}^n) \neq \emptyset$ (they contain the point $(y_{\bar{x}}, \bar{x})$) which implies that $\mathrm{ri}(S) = \mathrm{ri}(S_1) \cap \mathrm{ri}(S_2) \cap (\mathrm{ri}(Y) \times \mathbb{R}^n)$ and recalling the representations of $\mathrm{ri}(S_1)$ and $\mathrm{ri}(S_2)$, we see that $(y_{\bar{x}}, \bar{x})$ which satisfies (4.34) also belongs to $\mathrm{ri}(S)$, i.e., Slater condition (2.12) holds. Therefore, Proposition 2.3 provides a valid $2\varepsilon$-inexact cut for $\mathcal{Q}$.

Let us use the notation $\mathcal{C}_1(x) = \theta_1 + \langle \beta_1, x - \bar{x} \rangle$ and $\mathcal{C}_2(x) = \theta_2 + \langle \beta_2, x - \bar{x} \rangle$ for respectively the inexact cuts given by Propositions 2.3 and 4.1. In Proposition 4.2 below, we derive upper and lower bounds on $\theta_1 - \theta_2 = \mathcal{C}_1(\bar{x}) - \mathcal{C}_2(\bar{x})$ (observe that in the exact case, i.e., when $\varepsilon = 0$, clearly $\theta_1 = \theta_2$ and $\beta_1 = \beta_2$). This will be done using characterizations of $\varepsilon$-optimal feasible primal-dual solutions to obtain bounds for the terms $\langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle$ and $\max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y \rangle$ (which are clearly null if $\hat{y}$ and $(\hat{\lambda}, \hat{\mu})$ are optimal primal-dual solutions). In particular, we will show that $\langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle$ is between $-2\varepsilon$ and 0. To derive these bounds, we will assume that

(A0) the gradient of objective function $f(\cdot, \bar{x})$ (resp. of constraint function $g_i(\cdot, \bar{x})$) is $L_0$ (resp. $L_i$)-co-coercive with $L_i >, i = 0, \ldots, p$.

Recall that $F : \mathrm{Dom}(F) \subseteq \mathbb{R}^m \to \mathbb{R}^m$ is $L$-co-coercive on $\Omega \subseteq \mathrm{Dom}(F)$ if

$$L\langle y - x, F(y) - F(x) \rangle \geq \|F(y) - F(x)\|^2, \ \forall x, y \in \Omega.$$

Proposition 4.2. *Let the assumptions of Proposition 4.1 hold and assume that* $Y \times X \subset dom(g_i)$ *for all* $i = 1, \ldots, p$. *Take* $\bar{x} \in X$ *and let* $\mathcal{L}_{\bar{x}}$ *be any lower bound on* $\mathcal{Q}(\bar{x})$. *Let* $\mathcal{C}_1(x) = \theta_1 + \langle \beta_1, x - \bar{x} \rangle$ *and* $\mathcal{C}_2(x) = \theta_2 + \langle \beta_2, x - \bar{x} \rangle$ *be respectively the* *inexact cuts given by Propositions 2.3 and 4.1 taking $\varepsilon_D = \varepsilon_P = \varepsilon$. Assume that $f$ and $g_i, i = 1, \ldots, p$, satisfy (A0), that $Y$ is compact, and set*

$$\mathcal{U}_{\bar{x}} = \frac{f(y_{\bar{x}}, \bar{x}) - \mathcal{L}_{\bar{x}} + \varepsilon}{\min(-g_i(y_{\bar{x}}, \bar{x}), i = 1, \ldots, p)}, L = L_0 + \mathcal{U}_{\bar{x}} \max_{i=1,\ldots,p} L_i.$$

*Then we have*

$$-2\varepsilon \leq \mathcal{C}_1(\bar{x}) - \mathcal{C}_2(\bar{x}) \leq 2\varepsilon + 2D_Y\sqrt{L\varepsilon},$$

*where $D_Y$ is the diameter of $Y$.*

*Proof.* Recall that

$$\begin{array}{l} \mathcal{C}_1(\bar{x}) = f(\hat{y}, \bar{x}) - 2\varepsilon, \\ \mathcal{C}_2(\bar{x}) = f(\hat{y}, \bar{x}) + \langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle - \max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y \rangle, \end{array}$$

and that $(\hat{y}, \hat{\lambda}, \hat{\mu})$ satisfy

$$\hat{y} \in S(\bar{x}),\ \hat{\mu} \geq 0,\ f(\hat{y}, \bar{x}) \leq \mathcal{Q}(\bar{x}) + \varepsilon,\ \theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}) \geq \mathcal{Q}(\bar{x}) - \varepsilon, \tag{4.39}$$

where $S(x)$ is defined in (4.36) and $\theta_{\bar{x}}$ is the dual function given by (4.35).

By the subgradient inequality, if $L_x$ is the Lagrangian given in Proposition 4.1, we get

(4.40)

$$\begin{array}{rcl} \theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}) & = & \min_{y \in Y} L_{\bar{x}}(y, \hat{\lambda}, \hat{\mu}) \geq L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) + \min_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), y - \hat{y} \rangle \\ & = & f(\hat{y}, \bar{x}) + \langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle + \min_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), y - \hat{y} \rangle = \mathcal{C}_2(\bar{x}). \end{array}$$

Therefore,

$$\begin{array}{rcl} \mathcal{C}_1(\bar{x}) & = & f(\hat{y}, \bar{x}) - 2\varepsilon \\ & \geq & \theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}) - 2\varepsilon \text{ by weak duality,} \\ & \overset{(4.40)}{\geq} & \mathcal{C}_2(\bar{x}) - 2\varepsilon. \end{array} \tag{4.41}$$

We next provide an upper bound for $\mathcal{C}_1(\bar{x}) - \mathcal{C}_2(\bar{x})$. Indeed, (4.39) implies that

$$f(\hat{y}, \bar{x}) \leq \theta_{\bar{x}}(\hat{\lambda}, \hat{\mu}) + 2\varepsilon = \min_{y \in Y} \{L_{\bar{x}}(y, \hat{\lambda}, \hat{\mu}) : y \in Y\} + 2\varepsilon$$

and hence that

$$L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) = f(\hat{y}, \bar{x}) + \langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle \leq \min_{y \in Y} \{L_{\bar{x}}(y, \hat{\lambda}, \hat{\mu}) : y \in Y\} + \langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle + 2\varepsilon$$

where the first equality is due to $\hat{y} \in S(\bar{x})$. The last inequality in turn is equivalent to $\tilde{\varepsilon} := 2\varepsilon + \langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle$ satisfying

$$2\varepsilon \geq 2\varepsilon + \langle \hat{\mu}, g(\hat{y}, \bar{x}) \rangle = \tilde{\varepsilon} \geq 0, \quad 0 \in \partial_{\tilde{\varepsilon}} \Big( L_{\bar{x}}(\cdot, \hat{\lambda}, \hat{\mu}) + \delta_Y(\cdot) \Big) (\hat{y}) \tag{4.42}$$

where $\delta_Y(\cdot)$ is the indicator function of set $Y$ given by

$$\delta_Y(y) = \left\{ \begin{array}{ll} 0 & \text{if } y \in Y, \\ +\infty & \text{otherwise.} \end{array} \right.$$

It is easy to check that $\|\hat{\mu}\| \leq \mathcal{U}_{\bar{x}}$ (see for instance the proof of Proposition 2.3 in [15]) which easily implies that $L_{\bar{x}}(\cdot, \hat{\lambda}, \hat{\mu})$ is $L$-co-coercive (for the interested reader, we provide in Lemma 8.1 in the appendix the proof that a sum of $L_i$-co-coercive mappings $f_i$ is $(\sum_{i=1}^n L_i)$-co-coercive). Combining this observation with (4.42) and Lemma 3.2 in [32], we obtain that there exists $v$ satisfying:

$$v \in \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) + \partial_{\tilde{\varepsilon}} \delta_Y(\hat{y}), \quad \|v\| \leq \sqrt{2L\tilde{\varepsilon}} \overset{(4.42)}{\leq} 2\sqrt{L\varepsilon}. \tag{4.43}$$

It is well known that set $\partial_{\tilde{\varepsilon}} \delta_Y(\hat{y})$ is the $\tilde{\varepsilon}$-normal set to $Y$ at $\hat{y}$ given by

$$\partial_{\tilde{\varepsilon}} \delta_Y(\hat{y}) = \{z \in \mathbb{R}^m : \langle z, y - \hat{y}\rangle \leq \tilde{\varepsilon} \ \forall y \in Y\}$$

and therefore $v$ which satisfies (4.43) also satisfies

$$\langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) - v, \hat{y} - y\rangle \leq \tilde{\varepsilon}, \ \forall y \in Y \Leftrightarrow \max_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) - v, \hat{y} - y\rangle \leq \tilde{\varepsilon}. \tag{4.44}$$

We then obtain the following upper bound for $\mathcal{C}_1(\bar{x}) - \mathcal{C}_2(\bar{x})$:

$$\begin{array}{lcl}
\mathcal{C}_2(\bar{x}) & = & f(\hat{y}, \bar{x}) + \langle \hat{\mu}, g(\hat{y}, \bar{x})\rangle - \max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y\rangle \\
 & = & \mathcal{C}_1(\bar{x}) + 2\varepsilon + \langle \hat{\mu}, g(\hat{y}, \bar{x})\rangle - \max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y\rangle \\
 & \overset{(4.42)}{\geq} & \mathcal{C}_1(\bar{x}) - \max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y\rangle \\
 & \geq & \mathcal{C}_1(\bar{x}) - \max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}) - v, \hat{y} - y\rangle - \max\limits_{y \in Y} \langle v, \hat{y} - y\rangle \\
 & \overset{(4.44)}{\geq} & \mathcal{C}_1(\bar{x}) - \tilde{\varepsilon} - \|v\| \, D_Y \overset{(4.43)}{\geq} \mathcal{C}_1(\bar{x}) - 2\varepsilon - 2D_Y\sqrt{L\varepsilon},
\end{array}$$

which achieves the proof of the proposition. □

The upper and lower bounds on $\mathcal{C}_1(\bar{x}) - \mathcal{C}_2(\bar{x})$ given in Proposition 4.2 are continuous functions of $\varepsilon$ which go to 0 as $\varepsilon$ goes to 0. Also these bounds are respectively positive and negative for positive $\varepsilon$. This shows that they are both of good quality for small values of $\varepsilon$ and this analysis does not ensure that one of these two is always better (i.e., has a larger intercept at $\bar{x}$) than the other.

The analysis above (the proof of Proposition 4.2) is also interesting per-se since it offers ways of characterizing $\varepsilon$-optimal primal-dual solutions and allows us to derive bounds on the two quantities $\langle \hat{\mu}, g(\hat{y}, \bar{x})\rangle$ and $\max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y\rangle$ which, by the first order optimality conditions, are null if $\hat{y}$ and $(\hat{\lambda}, \hat{\mu})$ are respectively optimal primal and dual solutions. More precisely, if $\hat{y}$ (resp. $(\hat{\lambda}, \hat{\mu})$) is an $\varepsilon$-optimal feasible primal (resp. dual) solution, then we have shown that $-2\varepsilon \leq \langle \hat{\mu}, g(\hat{y}, \bar{x})\rangle \leq 0$ and $0 \leq \max\limits_{y \in Y} \langle \nabla_y L_{\bar{x}}(\hat{y}, \hat{\lambda}, \hat{\mu}), \hat{y} - y\rangle \leq 2D_Y\sqrt{L\varepsilon} + 2\varepsilon$.

## 5. ISDDP algorithm for nondifferentiable problems.

The objective of this section is to introduce and study new variants of ISDDP which use the inexact cuts built in the previous sections.

We consider multistage stochastic nonlinear optimization problems of the form (5.45)

$$\min_{x_1 \in X_1(x_0, \xi_1)} f_1(x_1, x_0, \xi_1) + \mathbb{E}\left[\min_{x_2 \in X_2(x_1, \xi_2)} f_2(x_2, x_1, \xi_2) + \mathbb{E}\left[\ldots \right.\right.$$
$$\left.\left. \ldots + \mathbb{E}\left[\min_{x_T \in X_T(x_{T-1}, \xi_T)} f_T(x_T, x_{T-1}, \xi_T)\right]\right]\right],$$

where $x_0$ is given, $(\xi_t)_{t=2}^T$ is a stochastic process, $\xi_1$ is deterministic, and

$$X_t(x_{t-1}, \xi_t) = \{x_t \in \mathbb{R}^n : A_t x_t + B_t x_{t-1} = b_t, g_t(x_t, x_{t-1}, \xi_t) \leq 0, x_t \in \mathcal{X}_t\}.$$

We make the following assumption on $(\xi_t)$:

(H) $(\xi_t)$ is interstage independent and for $t = 2, \ldots, T$, $\xi_t$ is a random vector taking values in $\mathbb{R}^K$ with a discrete distribution and a finite support $\Theta_t = \{\xi_{t1}, \ldots, \xi_{tN_t}\}$ with $p_{ti} = \mathbb{P}(\xi_t = \xi_{ti}) > 0, i = 1, \ldots, N_t$, while $\xi_1$ is deterministic.

In the sequel, we will denote by $A_{tj}$, $B_{tj}$, and $b_{tj}$ the realizations of $A_t, B_t$, and $b_t$ in $\xi_{tj}$.

For this problem, we can write Dynamic Programming equations: the first stage problem is

$$\mathcal{Q}_1(x_0) = \left\{ \begin{array}{l} \min_{x_1 \in \mathbb{R}^n} f_1(x_1, x_0, \xi_1) + \mathcal{Q}_2(x_1) \\ x_1 \in X_1(x_0, \xi_1) \end{array} \right. \tag{5.46}$$

for $x_0$ given and for $t = 2, \ldots, T$, $\mathcal{Q}_t(x_{t-1}) = \mathbb{E}_{\xi_t}[\mathfrak{Q}_t(x_{t-1}, \xi_t)]$ with

$$\mathfrak{Q}_t(x_{t-1}, \xi_t) = \left\{ \begin{array}{l} \min_{x_t \in \mathbb{R}^n} f_t(x_t, x_{t-1}, \xi_t) + \mathcal{Q}_{t+1}(x_t) \\ x_t \in X_t(x_{t-1}, \xi_t), \end{array} \right. \tag{5.47}$$

with the convention that $\mathcal{Q}_{T+1}$ is null.

We set $\mathcal{X}_0 = \{x_0\}$ and make the following assumptions (H1) on the problem data:

(H1): there exists $\varepsilon > 0$ such that for $t = 1, \ldots, T$,

1) $\mathcal{X}_t$ is a nonempty, compact, and convex set.
2) For every $j = 1, \ldots, N_t$, the function $f_t(\cdot, \cdot, \xi_{tj})$ is convex, proper, lower semicontinuous on $\mathcal{X}_t \times \mathcal{X}_{t-1}$ and for every $x_{t-1}$ $\mathcal{X}_{t-1}^{\varepsilon}$ we have
$$\mathcal{X}_t \subset \mathrm{dom}(f_t(\cdot, x_{t-1}, \xi_{tj})).$$
3) For every $j = 1, \ldots, N_t$, each component $g_{ti}(\cdot, \cdot, \xi_{tj}), i = 1, \ldots, p$, of function $g_t(\cdot, \cdot, \xi_{tj})$ is convex, lower semicontinuous and finite on $\mathcal{X}_t \times \mathcal{X}_{t-1}$.
4) $X_1(x_0, \xi_1) \neq \emptyset$ and for every $t = 2, \ldots, T$, for every $j = 1, \ldots, N_t$, for every $x_{t-1} \in \mathcal{X}_{t-1}^{\varepsilon}$, the set $\mathrm{ri}(\mathcal{X}_t) \cap X_t(x_{t-1}, \xi_{tj})$ is nonempty.
5) for every $t \geq 2$, for every $j = 1, \ldots, N_t$, there is $(x_{tj}, x_{t-1j}) \in \mathrm{ri}(\mathcal{X}_t) \times \mathcal{X}_{t-1}$ such that $g_t(x_{tj}, x_{t-1j}, \xi_{tj}) < 0$.

We are now in a position to describe the ISDDP algorithm for nondifferentiable optimization problems of form (5.45). The ISDDP algorithm given below combines SDDP with the inexact cuts derived in Section 2.2:

---

**ISDDP algorithm.**

---

Step 0) **Initialization.** Let $\mathcal{Q}_t^0 : \mathcal{X}_{t-1} \to \mathbb{R}$, $t = 2, \ldots, T+1$, be affine functions satisfying $\mathcal{Q}_t^0 \leq \mathcal{Q}_t$. Set $k = 1$.

Step 1) **Forward pass.** Setting $x_0^k = x_0$, generate a sample $(\tilde{\xi}_1^k, \tilde{\xi}_2^k, \ldots, \tilde{\xi}_T^k)$ from the distribution of $(\xi_1, \xi_2, \ldots, \xi_T)$ and for $t = 1, 2, \ldots, T$, compute a $\delta_t^k$-optimal solution $x_t^k$ of

$$\min \left\{ f_t(x_t, x_{t-1}^k, \tilde{\xi}_t^k) + \mathcal{Q}_{t+1}^{k-1}(x_t) : x_t \in X_t(x_{t-1}^k, \tilde{\xi}_t^k) \right\}. \tag{5.48}$$

Step 2) **Backward pass.**

```
        For $t = T, T-1, \ldots, 2$,
            For $j = 1, \ldots, N_t$,
                Compute an $\varepsilon_t^k$-optimal solution $x_{tj}^k$ of

$$\underline{\mathfrak{Q}}_t^k(x_{t-1}^k, \xi_{tj}) = \left\{ \begin{array}{ll} \min\limits_{x_t, z} \; f_t(x_t, z, \xi_{tj}) + \mathcal{Q}_{t+1}^k(x_t) & \\ A_{tj} x_t + B_{tj} z = b_{tj}, & \\ g_t(x_t, z, \xi_{tj}) \leq 0, & \\ x_t \in \mathcal{X}_t, & \\ z = x_{t-1}^k, & [\lambda_{tj}^k] \end{array} \right. \tag{5.49}$$

                and an $\varepsilon_t^k$-optimal dual solution $\lambda_{tj}^k$ of the dual of problem (5.49)
                obtained dualizing constraints $z = x_{t-1}^k$.
            End For
            Compute

                $\beta_t^k = \sum_{j=1}^{N_t} p_{tj} \lambda_{tj}^k$,
                $\theta_t^k = \sum_{j=1}^{N_t} p_{tj} \Big( f_t(x_{tj}^k, x_{t-1}^k, \xi_{tj}) + \mathcal{Q}_{t+1}^k(x_{tj}^k) - \langle \lambda_{tj}^k, x_{t-1}^k \rangle \Big)$

            and store the new cut

$$\mathcal{C}_t^k(x_{t-1}) := \theta_t^k - 2\varepsilon_t^k + \langle \beta_t^k, x_{t-1} \rangle$$

            for $\mathcal{Q}_t$, making up the new approximation $\mathcal{Q}_t^k = \max\{\mathcal{Q}_t^{k-1}, \mathcal{C}_t^k\}$.
        End For
Step 4) Do $k \leftarrow k+1$ and go to Step 1).
```

REMARK 5.1. *ISDDP algorithm given above applies both to differentiable and nondifferentiable problems. In the differentiable case (when all functions $f_t(\cdot, \cdot, \xi_{tj})$ and $g_{ti}(\cdot, \cdot, \xi_{tj})$ are differentiable), compared to ISDDP introduced in [15], the variant of ISDDP given above does not need to solve an additional optimization problem to obtain the intercept of the cut. However, all subproblems solved in the forward and backward passes have additional variables and constraints; the number of additional variables and constraints being the size of $x_{t-1}$ for stage $t$.*

When objective functions $f_t(\cdot, \cdot, \xi_{tj})$ have saddle point representations (which is the case of all "well structured" convex functions), we can also derive another variant of ISDDP that combines SDDP with the inexact cuts given in Section 3. For instance, assuming to alleviate notation that $f_t$ is deterministic of the form $f_t(x_t, x_{t-1})$ with saddle point representation

$$f_t(x_t, x_{t-1}) = x_{t-1}^T a_{t,1} + x_t^T a_{t,2} + \max_{w \in \mathcal{W}_t} x_t^T \bar{A}_t w + x_{t-1}^T \bar{B}_t w - \Psi_t(w), \tag{5.50}$$

setting

$$\begin{array}{l} \Delta_{k+1} = \{\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_k) \in \mathbb{R}^{k+1} : \lambda \geq 0, \sum_{i=0}^k \lambda_i = 1\}, \\ \bar{\theta}_t^{0:k} = [\theta_t^0; \theta_t^1 - 2\varepsilon_t^1; \ldots; \theta_t^k - 2\varepsilon_t^k], \; \beta_t^{0:k} = [\beta_t^0, \beta_t^1, \ldots, \beta_t^k], \\ \phi_{tk}(\lambda) = -\lambda^T \bar{\theta}^{0:k}, \end{array}$$

from the saddle point representation

$$\mathcal{Q}_{t+1}^k(x_t) = \max_{\lambda \in \Delta_{k+1}} \sum_{i=0}^k \lambda_i (\theta_t^i - 2\varepsilon_t^i + \langle \beta_t^i, x_t \rangle) = \max_{\lambda_2 \in \Delta_{k+1}} x_t^T \beta_t^{0:k} \lambda_2 - \phi_{t,k}(\lambda_2),$$

of $\mathcal{Q}_{t+1}^k$ where $\varepsilon_t^0 = 0$, we deduce the saddle point representation

$$x_{t-1}^T a_{t,1} + x_t^T a_{t,2} + \max_{\lambda \in \Lambda} x_t^T \mathcal{A}_t^k \lambda + x_{t-1}^T \mathcal{B}_t \lambda - \tilde{\phi}_{t,k}(\lambda) \tag{5.51}$$

of $f_t(x_t, x_{t-1}) + \mathcal{Q}_{t+1}^k(x_t)$ where

$$\begin{array}{l} \mathcal{A}_t^k = [\bar{A}_t, \beta_t^{0:k}],\ \mathcal{B}_t = [\bar{B}_t, 0],\ \tilde{\phi}_{t,k}(\lambda_1, \lambda_2) = \Psi_t(\lambda_1) + \phi_{t,k}(\lambda_2), \\ \Lambda = \{\lambda = (\lambda_1, \lambda_2) : \lambda_1 \in \mathcal{W}_t, \lambda_2 \in \Delta_{k+1}\}. \end{array}$$

In this situation, (5.51) provides a saddle point representation of the objective functions of problems (5.49) solved in the backward passes which allows us to build, using Section 3, inexact cuts of controlled accuracy for value functions $\underline{\mathfrak{Q}}_t^k(\cdot, \xi_{tj})$ and therefore for $\mathcal{Q}_t$.

We now study the convergence of ISDDP and start introducing more notation. Due to Assumption (H), the realizations of $(\xi_t)_{t=1}^T$ form a scenario tree of depth $T+1$ where the root node $n_0$ associated to a stage 0 (with decision $x_0$ taken at that node) has one child node $n_1$ associated to the first stage (with $\xi_1$ deterministic). We denote by $\mathcal{N}$ the set of nodes and for a node $n$ of the tree, we define:

- $C(n)$: the set of children nodes (the empty set for the leaves);
- $x_n$: a decision taken at that node;
- $p_n$: the transition probability from the parent node of $n$ to $n$;
- $\xi_n$: the realization of process $(\xi_t)$ at node $n$: for a node $n$ of stage $t$, this realization $\xi_n$ contains in particular the realizations $b_n$ of $b_t$, $A_n$ of $A_t$, and $B_n$ of $B_t$.

Next, we define for iteration $k$ decisions $x_n^k$ for all node $n$ of the scenario tree simulating the policy obtained in the end of iteration $k-1$ replacing cost-to-go function $\mathcal{Q}_t$ by $\mathcal{Q}_t^{k-1}$ for $t = 2, \ldots, T+1$:

**Simulation of ISDDP policy in the end of iteration $k-1$.**

```
For $t = 1, \ldots, T$,
  For every node $n$ of stage $t-1$,
      For every child node $m$ of node $n$, compute a $\delta_t^k$-optimal solution $x_m^k$ of
(5.52)   $\underline{\mathfrak{Q}}_t^{k-1}(x_n^k, \xi_m) = \left\{ \begin{array}{l} \inf_{x_m} \ f_t(x_m, x_n^k, \xi_m) + \mathcal{Q}_{t+1}^{k-1}(x_m) \\ A_m x_m + B_m x_n^k = b_m, \\ g_t(x_m, x_n^k, \xi_m) \leq 0, \\ x_m \in \mathcal{X}_t, \end{array} \right.$
          where $x_{n_0}^k = x_0$.
      End For
  End For
End For
```

We will assume that the sampling procedure in ISDDP satisfies the following property:

(H2) The samples in the backward passes are independent: $(\tilde{\xi}_2^k, \ldots, \tilde{\xi}_T^k)$ is a realization of $\xi^k = (\xi_2^k, \ldots, \xi_T^k) \sim (\xi_2, \ldots, \xi_T)$ and $\xi^1, \xi^2, \ldots,$ are independent.

As said in the introduction, a useful tool for the convergence analysis of SDDP and ISDDP is Lemma 5.2 in [11] for vanishing errors and Lemma 4.1 in [15] for bounded errors. We provide different proofs of these lemmas with slightly different assumptions, one of them being stronger (the continuity of $f$ [which is satisfied when the lemmas are applied to study the convergence of ISDDP]) and two being weaker. More precisely, in these lemmas we do not assume $f^n \leq f$ and take equicontinuous sequences $f^n$ instead of sequences of Lipschitz continuous functions. If we assumed $f^n \leq f$, the proof would be a little shorter, because boundedness of $\{f^n\}$ would be immediate. From these assumptions, we also derive a stronger conclusion, used in the convergence analysis.

Lemma 5.1. *Let $(X, d)$ be a compact metric space. If $\{x_n\}_{n\in\mathbb{N}}$ is a sequence in $X$, $\{f^n\}_{n\in\mathbb{N}}$ is an equicontinuous sequence of real functions on $X$, $f^1 \leq f^2 \leq f^3 \leq \ldots,$ and $f$ is a continuous real function on $X$ then the following conditions are equivalent:*
*(a)* $\lim_{m,n\to\infty} f^m(x_n) - f(x_n) = 0$.
*(b)* $\lim_{n\to\infty} f^n(x_n) - f(x_n) = 0$.
*Morever, if (a) or (b) holds then $f^n$ converges uniformly to a continuous function which coincides with $f$ on the set*

$$Y_* = \left\{ y \in X \ : \ y = \lim_{j\to\infty} x_{n_j} \textit{ for some subsequence } \{x_{n_j}\}_{j\in\mathbb{N}} \right\}.$$

*Proof.* See the Appendix. □

The proof of the previous lemma can be adapted to prove Lemma 5.2 which will be used in the convergence analysis of ISDDP with bounded errors.

Lemma 5.2. *Let $(X, d)$ be a compact metric space, let $f : X \to \mathbb{R}$ be continuous and suppose that the sequence of equicontinuous functions $f^k, k \in \mathbb{N}$ satisfies $f^k(x) \leq f^{k+1}(x)$ for all $x \in X,\ k \in \mathbb{N}$. Let $(x^k)_{k\in\mathbb{N}}$ be a sequence in $X$ and assume that*

$$\overline{\lim_{k\to+\infty}} f(x^k) - f^k(x^k) \leq S \tag{5.53}$$

*for some finite $S \geq 0$. Then*

$$\overline{\lim_{k\to+\infty}} f(x^k) - f^{k-1}(x^k) \leq S. \tag{5.54}$$

*Moreover, $f^n$ converges uniformly to a continuous function $g$ such that $|f(y)-g(y)| \leq S$ for every $y$ in the set*

$$Y_* = \left\{ y \in X \ : \ y = \lim_{j\to\infty} x_{n_j} \textit{ for some subsequence } \{x_{n_j}\}_{j\in\mathbb{N}} \right\}.$$

*Proof.* See the Appendix. □

We are now in a position to state our first convergence theorem for ISDDP.

Theorem 5.3 (Convergence of ISDDP with bounded errors). *Consider the sequences of decisions $(x_n^k)_{n\in\mathcal{N}}$ and of functions $(\mathcal{Q}_t^k)$ generated in the simulation of ISDDP. Assume that (H), (H1), and (H2) hold, and that errors $\varepsilon_t^k$ and $\delta_t^k$ are bounded: $0 \leq \varepsilon_t^k \leq \bar{\varepsilon}$, $0 \leq \delta_t^k \leq \bar{\delta}$ for finite $\bar{\delta}, \bar{\varepsilon}$. Then the following holds:*

*(i) for $t = 2, \ldots, T+1$, for all node $n$ of stage $t-1$, almost surely*

(5.55)
$$0 \leq \underline{\lim}_{k \to +\infty} \mathcal{Q}_t(x_n^k) - \mathcal{Q}_t^k(x_n^k) \leq \overline{\lim}_{k \to +\infty} \mathcal{Q}_t(x_n^k) - \mathcal{Q}_t^k(x_n^k) \leq (\bar{\delta} + 2\bar{\varepsilon})(T - t + 1);$$

*(ii) for every $t = 2, \ldots, T$, for all node $n$ of stage $t-1$, the limit superior and limit inferior of the sequence of upper bounds $\Big( \sum_{m \in C(n)} p_m (f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}(x_m^k)) \Big)_k$ satisfy almost surely*

(5.56)
$$\begin{array}{l} 0 \leq \underline{\lim}_{k \to +\infty} \displaystyle\sum_{m \in C(n)} p_m \Big[ f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}(x_m^k) \Big] - \mathcal{Q}_t(x_n^k), \\ \overline{\lim}_{k \to +\infty} \displaystyle\sum_{m \in C(n)} p_m \Big[ f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}(x_m^k) \Big] - \mathcal{Q}_t(x_n^k) \leq (\bar{\delta} + 2\bar{\varepsilon})(T - t + 1); \end{array}$$

*(iii) the limit superior and limit inferior of the sequence $\underline{\mathfrak{Q}}_1^{k-1}(x_0, \xi_1)$ of lower bounds on the optimal value $\mathcal{Q}_1(x_0)$ of (5.45) satisfy almost surely*

(5.57)
$$\mathcal{Q}_1(x_0) - \bar{\delta} T - 2\bar{\varepsilon}(T-1) \leq \underline{\lim}_{k \to +\infty} \underline{\mathfrak{Q}}_1^{k-1}(x_0, \xi_1) \leq \overline{\lim}_{k \to +\infty} \underline{\mathfrak{Q}}_1^{k-1}(x_0, \xi_1) \leq \mathcal{Q}_1(x_0);$$

*(iv) for $t = 2, \ldots, T$, almost surely the sequence of functions $(\mathcal{Q}_t^k)_k$ converges uniformly to a continuous function $\mathcal{Q}_t^*$ which is at most at distance $(\bar{\delta} + 2\bar{\varepsilon})(T - t + 1)$ from $\mathcal{Q}_t$ on every accumulation point $\bar{x}_n$ of the sequences $(x_n^k)_k$ for every node $n$ of stage $t-1$.*

*Proof.* (i) We show (5.55) for $t = 2, \ldots, T+1$, and all node $n$ of stage $t-1$ by backward induction on $t$. The relation holds for $t = T+1$. Now assume that it holds for $t+1$ for some $t \in \{2, \ldots, T\}$. Let us show that it holds for $t$. Take a node $n$ of stage $t-1$. Let $\mathcal{S}_n$ be the iterations where the sampled scenario passes through node $n$ and take an iteration $k \in \mathcal{S}_n$. It was shown in Lemma 5.2 in [15] that for the classes of problems we consider, Assumptions (H1)-3),5) imply that almost surely for every $j, k$, there exists $x_t$ satisfying

$$x_t \in \mathrm{ri}(\mathcal{X}_t), A_{tj} x_t + B_{tj} x_{t-1}^k = b_{tj} \text{ and } g_t(x_t, x_{t-1}^k, \xi_{tj}) < 0.$$

Recalling that $\mathcal{X}_t \times \mathcal{X}_{t-1} \subset \mathrm{dom}(g_{ti})$ for all $i$, we can reproduce the reasoning used just after Proposition 4.1 in Section 4 to deduce that for every $j, t$, there exists

$$(x_t, z) \in \mathrm{ri}(S_{tj}) \qquad (5.58)$$

where

$$S_{tj} = \{(x_t, z) : A_{tj} x_t + B_{tj} z = b_{tj}, g_t(x_t, z, \xi_{tj}) \leq 0, x_t \in \mathcal{X}_t\}.$$

Condition (5.58) is exactly Slater condition (2.12) (from Proposition 2.3) written for problem (5.49) solved in the backward pass of iteration $k$ for scenario $j$. Therefore, we can apply Proposition 2.3 to value function $\underline{\mathfrak{Q}}_t^k(\cdot, \xi_{tj})$ to obtain a $2\varepsilon_t^k$-inexact cut for this function for stage $t$ and iteration $k$ of ISDDP. More precisely, fix $j \in \{1, \ldots, N_t\}$ and take $m$ such that $\xi_{tj} = \xi_m$. Recalling that $\lambda_m^k$ is defined in (5.52) and setting

$$\mathcal{C}_{tm}^k(x_n) = f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}^k(x_n^k) - 2\varepsilon_t^k + \langle \lambda_m^k, x_n - x_n^k \rangle,$$

using Proposition 2.3, we get for all $x_n \in \mathcal{X}_{t-1}$ and $k \in \mathcal{S}_n$:

$$\mathcal{C}_{tm}^k(x_n) \leq \underline{\mathfrak{Q}}_t^k(x_n, \xi_m) \qquad (5.59)$$

and

$$\underline{\mathfrak{Q}}_t^k(x_n^k, \xi_m) - \mathcal{C}_{tm}^k(x_n^k) \leq 2\varepsilon_t^k. \tag{5.60}$$

This implies that $\mathcal{Q}_t^k$ is indeed a valid cut for $\mathcal{Q}_t$: for $x_n \in \mathcal{X}_{t-1}$ and $k \in \mathcal{S}_n$, we have

$$\begin{array}{lll}\mathcal{Q}_t(x_n) = \sum_{m \in C(n)} p_m \mathfrak{Q}_t(x_n, \xi_m) & \geq & \sum_{m \in C(n)} p_m \underline{\mathfrak{Q}}_t^k(x_n, \xi_m) \\ & \overset{(5.59)}{\geq} & \sum_{m \in C(n)} p_m \mathcal{C}_{tm}^k(x_n) = \mathcal{C}_t^k(x_n).\end{array} \tag{5.61}$$

Also by definition of $x_m^k$ computed in the simulation of iteration $k$ we get

$$f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}^{k-1}(x_m^k) \leq \underline{\mathfrak{Q}}_t^{k-1}(x_n^k, \xi_m) + \delta_t^k. \tag{5.62}$$

Therefore, for $k \in \mathcal{S}_n$:

$$\begin{array}{lll}\mathcal{C}_t^k(x_n^k) & = & \displaystyle\sum_{m \in C(n)} p_m \mathcal{C}_{tm}^k(x_n^k), \\ & \overset{(5.60)}{\geq} & \displaystyle\sum_{m \in C(n)} p_m \left[\underline{\mathfrak{Q}}_t^k(x_n^k, \xi_m) - 2\varepsilon_t^k\right], \\ & \geq & \displaystyle -2\bar{\varepsilon} + \sum_{m \in C(n)} p_m \underline{\mathfrak{Q}}_t^{k-1}(x_n^k, \xi_m), \\ & \overset{(5.62)}{\geq} & \displaystyle -2\bar{\varepsilon} + \sum_{m \in C(n)} p_m \left[f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}^{k-1}(x_m^k) - \delta_t^k\right], \\ & \geq & \displaystyle -2\bar{\varepsilon} - \bar{\delta} + \sum_{m \in C(n)} p_m \left[f_t(x_m^k, x_n^k, \xi_m) + \mathcal{Q}_{t+1}^{k-1}(x_m^k)\right].\end{array} \tag{5.63}$$

It follows that for $k \in \mathcal{S}_n$

$$\begin{array}{lll} 0 & \overset{(5.61)}{\leq} & \mathcal{Q}_t(x_n^k) - \mathcal{Q}_t^k(x_n^k) \leq \mathcal{Q}_t(x_n^k) - \mathcal{C}_t^k(x_n^k) \\ & \overset{(5.63)}{\leq} & \displaystyle 2\bar{\varepsilon} + \bar{\delta} + \sum_{m \in C(n)} p_m \left[\mathfrak{Q}_t(x_n^k, \xi_m) - f_t(x_m^k, x_n^k, \xi_m) - \mathcal{Q}_{t+1}^{k-1}(x_m^k)\right] \\ & \leq & \displaystyle 2\bar{\varepsilon} + \bar{\delta} + \sum_{m \in C(n)} p_m \Big[\underbrace{\mathfrak{Q}_t(x_n^k, \xi_m) - f_t(x_m^k, x_n^k, \xi_m) - \mathcal{Q}_{t+1}(x_m^k)}_{\leq 0 \text{ by definition of } \mathfrak{Q}_t \text{ and } x_m^k}\Big] \\ & & \displaystyle + \sum_{m \in C(n)} p_m \Big[\mathcal{Q}_{t+1}(x_m^k) - \mathcal{Q}_{t+1}^{k-1}(x_m^k)\Big].\end{array} \tag{5.64}$$

Using the induction hypothesis, we have for every $m \in C(n)$ that $\overline{\lim}_{k \to +\infty} \mathcal{Q}_{t+1}(x_m^k) - \mathcal{Q}_{t+1}^k(x_m^k) \leq (\bar{\delta} + 2\bar{\varepsilon})(T - t)$. Following the proof of Lemma 4.2 in [16], we obtain that sequence $(\beta_t^k)_k$ is almost surely bounded and that functions $(\mathcal{Q}_t^k)_k$ are $L$-Lipschitz continuous and therefore sequence $(\mathcal{Q}_t^k)_k$ is monotone and equicontinuous. Since $\mathcal{Q}_t$ is continuous on $\mathcal{X}_{t-1}$, we can apply Lemma 5.2 to obtain $\overline{\lim}_{k \to +\infty} \mathcal{Q}_{t+1}(x_m^k) - \mathcal{Q}_{t+1}^{k-1}(x_m^k) \leq (\bar{\delta} + 2\bar{\varepsilon})(T - t)$, which, plugged into (5.64), gives

$$\overline{\lim_{k \to +\infty, k \in \mathcal{S}_n}} \mathcal{Q}_t(x_n^k) - \mathcal{Q}_t^k(x_n^k) \leq (\bar{\delta} + 2\bar{\varepsilon})(T - t + 1). \tag{5.65}$$

Finally, to conclude the proof of (i), it remains to show that

$$\overline{\lim_{k \to +\infty, k \notin \mathcal{S}_n}} \mathcal{Q}_t(x_n^k) - \mathcal{Q}_t^k(x_n^k) \leq (\bar{\delta} + 2\bar{\varepsilon})(T - t + 1), \tag{5.66}$$

and with relation (5.65) at hand, relation (5.66) can be shown by contradiction following the end of the proof of Theorem 4.2 in [15].

(ii) and (iii) can be shown using (i) and following the proof of Theorem 4.2-(ii), (iii) in [15].

(iv) is an immediate consequence of (i) and Lemma 5.2. □

We can now state our second convergence theorem for ISDDP:

THEOREM 5.4 (Convergence of ISDDP with vanishing errors). *Consider the sequences of decisions $(x_n^k)_{n \in \mathcal{N}}$ and of functions $(\mathcal{Q}_t^k)$ generated in the simulation of ISDDP. Assume that (H), (H1), and (H2) hold, and that for all $t$ we have $\lim_{k \to +\infty} \varepsilon_t^k = \lim_{k \to +\infty} \delta_t^k = 0$. Then almost surely the limit of the sequence $(\underline{\mathfrak{Q}}_1^{k-1}(x_0, \xi_1))_{k \geq 1}$ is the optimal value $\mathcal{Q}_1(x_0)$ of (5.45). Moreover, for $t = 2, \ldots, T$, almost surely the sequence of functions $(\mathcal{Q}_t^k)_k$ converges uniformly to a continuous function $\mathcal{Q}_t^*$ which coincides with $\mathcal{Q}_t$ on every accumulation point $\bar{x}_n$ of the sequences $(x_n^k)_k$ for every node $n$ of stage $t-1$.*

*Proof.* It suffices to follow the proof of Theorem 5.3 and to use Lemma 5.1 instead of Lemma 5.2. □

If instead of the inexact cuts from Section 2 we use in ISDDP the inexact cuts from Section 3 based on saddle point representations of the objective, we obtain similar convergence results, due to the fact that the error terms in both the cuts from Section 2 and from Section 3 linearly depend on $\delta_t^k$ and $\varepsilon_t^k$.

## 6. Numerical experiments.

We consider the multistage nondifferentiable nonlinear stochastic program given by the following DP equations: the Bellman function for stage $t = 1, \ldots, T$, is $\mathcal{Q}_t(x_{t-1}) = \mathbb{E}_{\xi_t, \Psi_t, U_t}[\mathfrak{Q}_t(x_{t-1}, \xi_t, \Psi_t, U_t)]$ and for $t = 1, \ldots, T$, $\mathfrak{Q}_t(x_{t-1}, \xi_t, \Psi_t, U_t)$ is given by

$$
\begin{array}{l}
\min \; f_t(x_t, x_{t-1}, \xi_t, U_t) + \mathcal{Q}_{t+1}(x_t) \\
-100\,\mathbf{e} \leq x_t \leq 100\,\mathbf{e}, \\
\max(4(x_t - \mathbf{e})^T(x_t - \mathbf{e}), x_t^T(\xi_t \xi_t^T + \alpha I_n)x_t + x_t^T \xi_t + 1) \leq \Psi_t,
\end{array}
\tag{6.67}
$$

where $x_t \in \mathbb{R}^n$, $f_t(x_t, x_{t-1}, \xi_t, U_t) = \max((x_t - x_{t-1})^T(\xi_t \xi_t^T + \alpha I_n)(x_t - x_{t-1}) + x_t^T \xi_t + 1, x_t^T(\xi_t \xi_t^T + \alpha I_n)x_t + x_t^T \mathbf{e} + U_t)$, $\mathbf{e}$ is a vector of size $n$ of ones, and $\mathcal{Q}_{T+1}$ is the null function. In these equations, $\alpha \geq 0$ is a parameter, $\xi_t$ is a discretization of a Gaussian random vector with mean vector $m_t$ having entries 1 or $-1$ and covariance matrix $\Sigma_t = A_t A_t^T + 0.5I$ where $A_t$ has entries in $[-0.5, 0.5]$; $U_t$ is a discrete random variable taking values $+10$, $-10$, and $\Psi_t$ has discrete distribution with support contained in $[10^4, 10^5]$. The number of realizations $N_t$ for $(\xi_t, \Psi_t, U_t)$ is fixed to $N_t = N$ for each stage. We assume that $(\xi_1, \Psi_1, U_1)$ is known and $(\xi_2, \Psi_2, U_2), \ldots, (\xi_T, \Psi_T, U_T)$ are independent.

We generate 2 instances of this problem with parameters $\alpha = 0.2$ and $T, n, M$ given by $(T, n, M) = (5, 10, 20)$ and $(T, n, M) = (5, 50, 20)$. The instances are chosen taking realizations $\Psi_{tj}$ of $\Psi_t$ sufficiently large, in such a way that Assumption (H1)-4) holds. It is easy to check that the remaining assumptions (H1) and (H) are satisfied and therefore SDDP can be applied to solve the problem as well as SDDP combined with the inexact cuts from Section 2. In what follows, we denote the corresponding solution methods by `SDDP` and `ISDDPND` (Inexact SDDP for nondifferentiable problems). We also solved problem (6.67) using Stochastic Dynamic Cutting Plane (denoted by `StoDCuP`), StoDCuP combined with inexact cuts (denoted by `IStoDCuP`) introduced in [16] as well as the inexact variant of SDDP introduced in [15] that we

| Iteration | 1-10 | 11-20 | 21-40 | 41-140 | 141-240 | 241-350 | > 350 |
|---|---|---|---|---|---|---|---|
| Parameter value | 10 | 5 | 3 | 1 | 0.5 | 0.1 | e-6 |

TABLE 1

*Relative error of the subproblem solutions along iterations of inexact methods (Mosek parameter MSK_DPAR_INTPNT_TOL_REL_GAP).*

| | IMSDDP | ISDDPND | ISDDPD | IStoDCuP | MSDDP | SDDP | StoDCuP |
|---|---|---|---|---|---|---|---|
| Iterations | 439 | 409 | 465 | 655 | 569 | 431 | 770 |
| CPU time | 233.1 | 282.2 | 322.5 | 582.4 | 352.7 | 297.3 | 791.8 |

$T = 5, n = 10, M = 20$

| | IMSDDP | ISDDPND | ISDDPD | IStoDCuP | MSDDP | SDDP | StoDCuP |
|---|---|---|---|---|---|---|---|
| Iterations | 400 | 400 | 400 | - | 400 | 400 | - |
| CPU time | 3 424 | 4 387 | 3 237 | - | 3 547 | 4 504 | - |

$T = 5, n = 50, M = 20, \alpha = 0.2$

TABLE 2

*Number of iterations and CPU time in seconds needed to solve the two instances. For the second instance, the unfilled cells for* IStoDCuP *and* StoDCuP *indicate that these methods had not converged after completing the maximal number of 600 iterations. It took* IStoDCuP *(resp.* StoDCuP*) 2 230 s. (resp. 2 356 s.) to complete these 600 iterations.*

will denote by ISDDPD (Inexact SDDP for differentiable problems) in what follows (the interested reader can find in the Appendix the formulas for the inexact cuts to use for this inexact variant of SDDP). Observe that this inexact variant ISDDPD was designed for differentiable problems but can be applied to (6.67) reformulating the problem as a differentiable problem replacing in (6.67) each max with 2 quadratic constraints. Finally, we consider a mixed StoDCuP-SDDP variant (denoted by MSDDP) which uses StoDCuP for the first 150 iterations and SDDP for the remaining iterations, as well as its inexact counterpart (denoted by IMSDDP) which is StoDCuP with inexact cuts, i.e., IStoDCuP, for the first 150 iterations and SDDP with the inexact cuts from Section 2, i.e., ISDDPND, for the remaining iterations. The Matlab implementation of all methods can be found at https://github.com/vguigues/ISDDP_NLP. All subproblems were solved using Mosek optimization library [1].

For the inexact variants with inexact cuts to be well defined, we also need to set the level of accuracy of the computed solutions along the iterations of the methods. In our experiments, the relative error of the subproblem solutions (Mosek parameter MSK_DPAR_INTPNT_TOL_REL_GAP whose range is any value $\geq 10^{-14}$ and default value is $10^{-8}$) is given in Table 1; see also Remark 2 in [15] for other choices of sequences of noises $\varepsilon_t^k$. For the exact variants, this parameter was set to $10^{-10}$ for all iterations.

All methods compute at each iteration a lower bound on the optimal value which is the optimal value of the first stage problem solved in the forward pass and upper bounds computed by Monte-Carlo simulations, from iteration 400 on, using the last 400 forward scenarios. The algorithms stopped when a relative gap of at most 0.1 was achieved or, for the largest instance, when the maximal number of 600 iterations was reached.

The number of iterations before stopping the algorithms as well as the CPU time is given in Table 2 for all methods and the two instances.

The evolution of the upper and lower bounds for some iterations, all methods, and the two instances is given in Tables 3 and 4.

| Iteration | IMSDDP | ISDDPND | ISDDPD | IStoDCuP | MSDDP | SDDP | StoDCuP |
|---|---|---|---|---|---|---|---|
| 400 | 14.34 | 14.66 | 14.32 | 5.07 | 14.35 | 14.66 | 2.76 |
| 409 | 14.39 | 14.66 | 14.41 | 6.07 | 14.41 | 14.66 | 4.67 |
| 431 | 14.46 | - | 14.45 | 9.17 | 14.46 | 14.67 | 7.47 |
| 439 | 14.48 | - | 14.49 | 9.45 | 14.49 | - | 8.95 |
| 465 | - | - | 14.62 | 12.80 | 14.57 | - | 12.34 |
| 500 | - | - | - | 12.80 | 14.57 | - | 12.34 |
| 569 | - | - | - | 13.71 | 14.62 | - | 13.57 |
| 770 | - | - | - | - | - | - | 13.97 |

$T = 5, n = 10, M = 20, \alpha = 0.2$

| Iteration | IMSDDP | ISDDPND | ISDDPD | IStoDCuP | MSDDP | SDDP | StoDCuP |
|---|---|---|---|---|---|---|---|
| 200 | -96 077 | 84.8 | 84.4 | -1.832e6 | -8 884 | 83.8 | -1.843e6 |
| 300 | 53.7 | 85.8 | 85.7 | -1.05e6 | 35.1 | 85.6 | -1.0e6 |
| 400 | 84.6 | 85.9 | 85.9 | -6.6e5 | 84.5 | 85.9 | -7.2e5 |
| 600 | - | - | - | -3.3e4 | - | - | -3.5e4 |

$T = 5, n = 50, M = 20, \alpha = 0.2$

Table 3

*Lower bounds computed along the iterations of the methods for both instances.*

We observe that the sequences of upper bounds decrease and as expected the sequences of lower bounds are increasing and both sequences converge to the same values. On these instances, `StoDCuP` and its inexact variant `IStoDCuP` need more iterations and time than the other methods to converge (for the largest instance the maximal number of 600 iterations was even not enough for `StoDCuP` and `IStoDCuP` to converge). However, we observed that the first iterations of `StoDCuP` and `IStoDCuP` are much quicker than the first iterations of `SDDP` and its inexact variants, which explains the good performance of the mixed StoDCuP-SDDP method and its inexact counterpart. Indeed, `IMSDDP` is the quickest to converge for the first instance and the second quickest, after `ISDDPD`, for the largest instance. In particular, both `MSDDP` and `IMSDDP` converge much quicker than `SDDP`. Out of the 8 runs of the inexact methods, only one did not converge quicker than its exact counterpart, namely `ISSDPD` for the smallest instance. Among inexact variants `ISSDPD` and `ISSDPND` of `SDDP`, method `ISSDPD` was the quickest on the instance with the largest value of the state vector size $n$ ($n = 50$) while `ISSDPND` was the quickest on the smallest instance, which may come from the increase in the CPU time needed to solve subproblems with `ISSDPND` due to the copy of variables used to derive the cuts. On the other hand, `ISSDPND` is more general and can apply to nondifferentiable problems contrary to `ISSDPD`.

**7. Conclusion.** In [15], an inexact variant of SDDP called ISDDP was introduced. Two variants of the method were described in [15]: one for linear problems and one for nonlinear differentiable problems. In this paper, we explained how to extend ISDDP for nondifferentiable multistage stochastic programs. We provided formulas to compute inexact cuts for value functions of possibly nondifferentiable optimization problems and combined these cuts with SDDP to describe two new inexact variants of SDDP, one for each of the classes of cuts derived (the cuts from Section 2 and the cuts from Section 3).

Several comments are in order:

- the variants of ISDDP presented in this paper can be used both for nonlinear differentiable and nonlinear nondifferentiable optimization problems.
- For errors bounded from above by $\varepsilon$, same as ISDDP for linear programs

| Iteration | IMSDDP | ISDDPND | ISDDPD | IStoDCuP | MSDDP | SDDP | StoDCuP |
|---|---|---|---|---|---|---|---|
| 400 | 20.81 | 17.79 | 20.4 | 32.61 | 22.17 | 20.55 | 43.19 |
| 409 | 19.03 | 15.72 | 18.3 | 27.47 | 21.94 | 19.79 | 26.32 |
| 431 | 16.48 | - | 17.1 | 19.85 | 18.57 | 16.25 | 16.25 |
| 439 | 15.89 | - | 16.81 | 18.78 | 18.56 | - | 20.03 |
| 465 | - | - | 15.9 | 18.42 | 18.14 | - | 19.37 |
| 500 | - | - | - | 17.11 | 16.75 | - | 17.72 |
| 569 | - | - | - | 16.42 | 16.22 | - | 16.86 |
| 770 | - | - | - |  | - | - | 15.94 |

$T = 5, n = 10, M = 20, \alpha = 0.2$

| Iteration | IMSDDP | ISDDPND | ISDDPD | IStoDCuP | MSDDP | SDDP | StoDCuP |
|---|---|---|---|---|---|---|---|
| 400 | 86.22 | 86.7 | 86.1 | 21 348 | 87.7 | 89.0 | 19 538 |
| 600 | - | - | - | 9 342 | - | - | 7 231 |

$T = 5, n = 50, M = 20, \alpha = 0.2$

TABLE 4

*Upper bounds computed along the iterations of the methods for both instances..*

introduced in [15], ISDDP variants of this paper provide $3\varepsilon T$-optimal first stage solutions. Using the analysis of Section 4, it is easy to check that ISDDP for nonlinear stochastic programs from [15] provides for bounded errors a $O(T\sqrt{\varepsilon})$-optimal first stage solution. However, all subproblems solved in the forward and backward passes of the variant of ISDDP that uses the cuts from Section 2 have additional variables and constraints; the number of additional variables and constraints being the size of $x_{t-1}$ for stage $t$.

- All variants of ISDDP from [15] and from this paper converge to an optimal policy for vanishing noises. The convergence analysis of ISDDP applied to nonlinear programs in [15] was however more technical due to the fact that the error terms in the inexact cuts were not a linear function of $\delta_t^k$ and $\varepsilon_t^k$ (see Proposition 5.4 in [15]).

## 8. Appendix.

LEMMA 8.1. *Assume that $F_i : \mathbb{R}^m \to \mathbb{R}^m$ is $L_i$-co-coercive for $i = 1, \ldots, n$. Then $\sum_{i=1}^n F_i$ is $(\sum_{i=1}^n L_i)$-co-coercive.*

*Proof.* We can assume w.l.o.g that all $L_i$ are positive. Let $S(x) = \sum_{i=1}^n F_i(x)$, $L = \sum_{i=1}^n L_i > 0$ , and $\alpha_i = \frac{L_i}{L}$. Observing that $\sum_{i=1}^n \alpha_i = 1$ and using the convexity of $\| \cdot \|^2$ we get:

$$\begin{array}{lll} \langle y - x, S(y) - S(x) \rangle & \geq & \sum_{i=1}^n \frac{1}{L_i} \|F_i(x) - F_i(y)\|^2 \\ & = & \frac{1}{L} \sum_{i=1}^n \alpha_i \| \frac{1}{\alpha_i} (F_i(x) - F_i(y))\|^2 \\ & \geq & \frac{1}{L} \|S(y) - S(x)\|^2, \end{array} \tag{8.68}$$

which achieves the proof of the lemma. □

**Proof of Lemma 5.1.** Implication (a)⇒(b) holds trivially. Suppose (b) holds. Since $X$ is compact and $f$ is continuous, the sequence $\{f^n(x_n)\}$ is bounded. Combining this result with the compactness of $X$ and the equicontinuity of $\{f^n\}$ we conclude that this sequence is pointwise uniformly bounded. Hence the monotone sequence $\{f^n(x)\}$ converges for any $x \in X$. Recall that $Y_*$ is the set of limit points of $\{x_n\}$ and let $g : X \to \mathbb{R}$ be the pointwise limit of $\{f^n\}$ that is,

$$g(x) = \lim_{n \to \infty} f^n(x) \quad (x \in X).$$

We claim that

1. $g$ is continuous;
2. $\{f^n\}$ converges uniformly to $g$;
3. $g(y) = f(y)$ for any $y \in Y_*$.

Continuity of $g$ follows from the equicontinuity of $\{f^n\}$ and its convergence to $g$. Since $\{f^n\}$ is a sequence of equicontinuous functions converging monotonically in a compact set to a continuous function $g$, this convergence is uniform. To prove item 3, suppose that $\lim_{j\to\infty} x_{n_j} = y$. Direct use of the triangle inequality yields

$$|f^{n_j}(y) - f(y)| \leq |f^{n_j}(y) - f^{n_j}(x_{n_j})| + |f^{n_j}(x_{n_j}) - f(x_{n_j})| + |f(x_{n_j}) - f(y)|.$$

It follows from the equicontinuity of $\{f^n\}$, the continuity of $f$, and the convergence of $\{x_{n_j}\}$ to $y$ that the first and third terms in the right-hand side of the above inequality converge to 0, while it follows from Assumption (b) that the middle term also converges to 0. Since $\{f^{n_j}(y)\}$ converges to $g(y)$, we have $g(y) = f(y)$.

To end the proof, take $\varepsilon > 0$. There exists $M_0 \in \mathbb{N}$ such that

$$m \geq M_0 \Rightarrow |f^m(x) - g(x)| < \varepsilon \qquad \forall x \in X.$$

It follows from the continuity of $f$ and $g$, and from the compactness of $X$ that there is $\delta > 0$ such that

$$d(x, x') \leq \delta \Rightarrow \; |f(x) - f(x')| \leq \varepsilon, \; |g(x) - g(x')| \leq \varepsilon.$$

It follows from the definition of $Y_*$ and the compactness of $X$ that there is $N_0 \in \mathbb{N}$ such that $d(x^n, Y_*) < \delta$ for $n \geq N_0$. Suppose that $m \geq M_0$ and $n \geq N_0$. There is $y \in Y_*$ such that $d(x^n, y) < \delta$. Therefore

(8.69)
$$\begin{aligned} |f^m(x_n) - f(x_n)| &\leq |f^m(x_n) - g(x_n)| + |g(x_n) - g(y)| + |g(y) - f(x_n)| \\ &= |f^m(x_n) - g(x_n)| + |g(x_n) - g(y)| + |f(y) - f(x_n)| < 3\varepsilon, \end{aligned}$$

which achieves the proof of the lemma. □

**Proof of Lemma 5.2.** The proof is a simple extension of the proof of Lemma 5.1. We outline the changes in the proof below. Since the sequence $f^n(x_n) - f(x_n)$ is bounded from above and $f$ is continuous on the compact set $X$, the sequence $f^n(x_n)$ is bounded from above. Same as in Lemma 5.1, together with the equicontinuity, the monotonicity of $f^n$, and the compactness of $X$, this implies that the sequence $f^n(x)$ converges for every $x \in X$ uniformly to a continuous function $g$. For every $y \in Y_*$, taking $\{x_{n_j}\}$ satisfying $y = \lim_{j\to\infty} x_{n_j}$, we get

$$\begin{aligned} |g(y) - f(y)| &= |\lim_{j\to\infty} f^{n_j}(y) - f(\lim_{j\to\infty} x_{n_j})| = |\lim_{j\to\infty} f^{n_j}(y) - f(x_{n_j})| \\ &\leq |\lim_{j\to\infty} f^{n_j}(y) - f^{n_j}(x_{n_j})| + |\lim_{j\to\infty} f^{n_j}(x_{n_j}) - f(x_{n_j})| = S. \end{aligned}$$

To conclude, it suffices to modify the last inequality (8.69) in Lemma 5.1 by

$$\begin{aligned} &|f^m(x_n) - f(x_n)| \leq |f^m(x_n) - g(x_n)| + |g(x_n) - g(y)| + |g(y) - f(x_n)| \\ &\leq |f^m(x_n) - g(x_n)| + |g(x_n) - g(y)| + |g(y) - f(y)| + |f(y) - f(x_n)| \\ &\leq S + 3\varepsilon, \end{aligned}$$

which concludes the proof of the lemma. □

**Formulas for inexact cuts for ISDDP from [15] applied to problem (6.67).** The inexact cut for ISDDP from [15] applied to problem (6.67) for $\mathcal{Q}_t$ takes the form $\mathcal{C}_t^k(x_{t-1}) = \theta_t^k - \eta_t^k + \langle \beta_t^k, x_{t-1}\rangle$ for iteration $k$. This cut is computed as follows. Given trial point $x_{t-1}^k$ we compute for $j = 1, \ldots, N_t$, an approximate optimal primal-dual solution $(f_{tj}^*, q_{tj}^*, x_{tj}^*, \lambda_{1j}^*)$ of

$$\begin{array}{l}
\min_{f,q,x_t} \;\; f + q\\
f \geq (x_t - x_{t-1}^k)^T(\xi_{tj}\xi_{tj}^T + \alpha I_n)(x_t - x_{t-1}^k) + x_t^T \xi_{tj} + 1, \;\; [\lambda_{1j}]\\
f \geq x_t^T(\xi_{tj}\xi_{tj}^T + \alpha I_n)x_t + x_t^T \mathbf{e} + U_{tj},\\
4(x_t - \mathbf{e})^T(x_t - \mathbf{e}) \leq \Psi_{tj},\\
x_t^T(\xi_{tj}\xi_{tj}^T + \alpha I_n)x_t + x_t^T\xi_{tj} + 1 \leq \Psi_{tj},\\
-100\,\mathbf{e} \leq x_t \leq 100\,\mathbf{e},\\
q \geq \theta_{t+1}^i + \langle \beta_{t+1}^i, x_t\rangle - \eta_{t+1}^i, \; i = 0, \ldots, k,
\end{array} \tag{8.70}$$

where $\lambda_{1j}^*$ is an approximate value for the optimal Lagrange multiplier associated to the first constraint (any approximate primal-dual solution can be used, for instance running a few iterations of a quadratic solver). We then define the Lagrangian $L(f, q, x_t, x_{t-1}, \lambda_1, \xi_t) = f + q + \lambda_1((x_t - x_{t-1})^T(\xi_t\xi_t^T + \alpha I_n)(x_t - x_{t-1}) + x_t^T\xi_t + 1 - f)$ obtained dualizing the coupling constraint and compute for $j = 1, \ldots, N_t$, the optimal value $\eta_{tj}^k$ of

$$\begin{array}{l}
\displaystyle\min_{f,q,x_t} \; (1 - \lambda_{1j}^*)(f - f_{tj}^*) + \langle \lambda_{1j}^*(\xi_{tj} + 2(\xi_{tj}\xi_{tj}^T + \alpha I_n)(x_{tj}^* - x_{t-1}^k)), x_t - x_{tj}^*\rangle + q - q_{tj}^*\\
\bar{f}_{tj} \geq f \geq x_t^T(\xi_{tj}\xi_{tj}^T + \alpha I_n)x_t + x_t^T\mathbf{e} + U_{tj},\\
4(x_t - \mathbf{e})^T(x_t - \mathbf{e}) \leq \Psi_{tj},\\
x_t^T(\xi_{tj}\xi_{tj}^T + \alpha I_n)x_t + x_t^T\xi_{tj} + 1 \leq \Psi_{tj},\\
-100\,\mathbf{e} \leq x_t \leq 100\,\mathbf{e},\\
q \geq \theta_{t+1}^i + \langle \beta_{t+1}^i, x_t\rangle - \eta_{t+1}^i, \; i = 0, \ldots, k,
\end{array}$$

where $\bar{f}_{tj}$ is an upper bound for $f_t(\cdot, \cdot, \xi_{tj})$ on $\mathcal{X}_t \times \mathcal{X}_{t-1} := \{x_t : -100\,\mathbf{e} \leq x_t \leq 100\,\mathbf{e}\} \times \{x_{t-1} : -100\,\mathbf{e} \leq x_{t-1} \leq 100\,\mathbf{e}\}$. Setting $\beta_{tj}^k = 2\lambda_{1j}^*(\xi_{tj}\xi_{tj}^T + \alpha I_n)(x_{t-1}^k - x_{tj}^*)$ and

$$\theta_{tj}^k = L(f_{tj}^*, q_{tj}^*, x_{tj}^*, x_{t-1}^k, \lambda_{1j}^*, \xi_{tj}) - \langle \beta_{tj}^k, x_{t-1}^k\rangle,$$

the coefficients $\theta_t^k, \eta_t^k, \beta_t^k$ of the cut $\mathcal{C}_t^k$ are given by

$$\theta_t^k = \sum_{j=1}^{N_t} p_{tj}\theta_{tj}^k, \;\; \beta_t^k = \sum_{j=1}^{N_t} p_{tj}\beta_{tj}^k, \text{ and } \eta_t^k = \sum_{j=1}^{N_t} p_{tj}\eta_{tj}^k.$$

If instead of approximate primal-dual solutions we compute exact primal-dual solutions, we get $\eta_{tj}^k = 0$, $L(f_{tj}^*, q_{tj}^*, x_{tj}^*, x_{t-1}^k, \lambda_{1j}^*, \xi_{tj}) = f_{tj}^* + q_{tj}^*$ and we get the usual cut computed by SDDP applied to convex problems.

## REFERENCES

[1] E. D. Andersen and K.D. Andersen. *The MOSEK optimization toolbox for MATLAB manual. Version 9.2*, 2019. https://www.mosek.com/documentation/.

[2] M. Bandarra and V. Guigues. Single cut and multicut sddp with cut selection for multistage stochastic linear programs: convergence proof and numerical experiments. *Computational Management Science, to appear.* https://arxiv.org/abs/1902.06757.

[3] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization.* Princeton University Press, 2009.

[4] A. Ben-Tal, A. Goryashko, E. Guslitzer, and A. Nemirovski. Adjustable robust counterpart of uncertain linear programs. *Mathematical Programming*, 99:351–376, 2003.

[5] A. Ben-Tal and A. Nemirovski. *Lectures on Modern Convex Optimization.* MOS-SIAM Series on Optimization, 2001.

[6] J.F. Benders. Partitioning Procedures for Solving Mixed-Variables Programming Problems. *Nmer. Math.*, 4:238–252, 1962.

[7] J.R. Birge. Decomposition and partitioning methods for multistage stochastic linear programs. *Oper. Res.*, 33:989–1007, 1985.

[8] J.R. Birge and C. J. Donohue. The Abridged Nested Decomposition Method for Multistage Stochastic Linear Programs with Relatively Complete Recourse. *Algorithmic of Operations Research*, 1:20–30, 2001.

[9] Z.L. Chen and W.B. Powell. Convergent Cutting-Plane and Partial-Sampling Algorithm for Multistage Stochastic Linear Programs with Recourse. *J. Optim. Theory Appl.*, 102:497–524, 1999.

[10] L. Ding and A. Shapiro. Stationary multistage programs. *Optimization Online*, 2019. http://www.optimization-online.org/DB_HTML/2019/09/7367.html.

[11] P. Girardeau, V. Leclere, and A.B. Philpott. On the convergence of decomposition methods for multistage stochastic convex programs. *Mathematics of Operations Research*, 40:130–145, 2015.

[12] V. Guigues. Inexact Stochastic Mirror Descent for two-stage nonlinear stochastic programs. *Mathematical Programming, to appear.* https://arxiv.org/pdf/1805.11732.pdf.

[13] V. Guigues. SDDP for some interstage dependent risk-averse problems and application to hydro-thermal planning. *Computational Optimization and Applications*, 57:167–203, 2014.

[14] V. Guigues. Convergence analysis of sampling-based decomposition methods for risk-averse multistage stochastic convex programs. *SIAM Journal on Optimization*, 26:2468–2494, 2016.

[15] V. Guigues. Inexact cuts in Stochastic Dual Dynamic Programming. *Siam Journal on Optimization*, 30:407–438, 2020.

[16] V. Guigues and R. Monteiro. Stochastic Dynamic Cutting Plane for multistage stochastic convex programs. *Journal of Optimization Theory and Applications, to appear.* https://arxiv.org/abs/1912.11946.

[17] V. Guigues and W. Römisch. Sampling-based decomposition methods for multistage stochastic programs based on extended polyhedral risk measures. *SIAM J. Optim.*, 22:286–312, 2012.

[18] J.L. Higle and S. Sen. *Stochastic Decomposition.* Kluwer, Dordrecht, 1996.

[19] M. Hindsberger and A. B. Philpott. Resa: A method for solving multi-stage stochastic linear programs. *SPIX Stochastic Programming Symposium*, 2001.

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

[21] V. Kozmik and D.P. Morton. Evaluating policies in risk-averse multi-stage stochastic programming. *Mathematical Programming*, 152:275–300, 2015.

[22] D. Kuhn, W. Wiesemann, and A. Georghiou. Primal and dual linear decision rules in stochastic and robust optimization. *Mathematical Programming*, 130:177–209, 2011.

[23] R. P. Liu and A. Shapiro. Risk neutral reformulation approach to risk averse stochastc programming. *arXiv*, 2019. https://arxiv.org/abs/1901.01302.

[24] Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming Series*, 103:127–152, 2005.

[25] M.V.F. Pereira and L.M.V.G Pinto. Multi-stage stochastic optimization applied to energy planning. *Math. Program.*, 52:359–375, 1991.

[26] A. Philpott, J.F. Bonnans, and F. Wahid. Midas: A mixed integer dynamic approximation scheme. *Mathematical Programming, to appear.*

[27] A. B. Philpott and Z. Guan. On the convergence of stochastic dual dynamic programming and related methods. *Oper. Res. Lett.*, 36:450–455, 2008.

[28] W.P. Powell. *Approximate Dynamic Programming.* John Wiley and Sons, 2nd edition, 2011.

[29] A. Shapiro. Analysis of stochastic dual dynamic programming method. *European Journal of Operational Research*, 209:63–72, 2011.

[30] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, Philadelphia, 2009.

[31] Z. Shixuan and X.A. Sun. Stochastic dual dynamic programming for multistage stochastic mixed-integer nonlinear optimization. *arXiv:1912.13278*, 2019.

[32] H. Yunlong and R.D.C. Monteiro. Accelerating Block-Decomposition First-Order Methods for Solving Composite Saddle-Point and Two-Player Nash Equilibrium Problems. *SIAM Journal on Optimization*, 25:2182–2211, 2015.

[33] G. Zakeri, A.B. Philpott, and D.M. Ryan. Inexact Cuts in Benders Decomposition. *SIAM Journal on Optimization*, 10:643–657, 2000.