Mathematical Programming manuscript No.
(will be inserted by the editor)

A Regularized Smoothing Method for Fully Parameterized Convex Problems with Applications to Convex and Nonconvex Two-Stage Stochastic Programming

Pedro Borges · **Claudia Sagastizábal** · **Mikhail Solodov**

Received: January 2020 / Accepted: date

Abstract We present an approach to regularize and approximate solution mappings of parametric convex optimization problems that combines interior penalty (log-barrier) solutions with Tikhonov regularization. Because the regularized mappings are single-valued and smooth under reasonable conditions, they can be used to build a computationally practical smoothing for the associated optimal value function. The value function in question, while resulting from parameterized convex problems, need not be convex. One motivating application of interest is two-stage (possibly nonconvex) stochastic programming. We show that our approach, being computationally implementable, provides locally bounded upper bounds for the subdifferential of the value function of qualified convex problems. As a by-product of our development, we also recover that in the given setting the value function is locally Lipschitz continuous. Numerical experiments are presented for two-stage convex stochastic programming problems, comparing the approach with the bundle method for nonsmooth optimization.

keywords Smoothing Techniques, Interior Penalty Solutions, Tikhonov Regularization, Nonconvex Stochastic Programming, Two-Stage Stochastic Programming, Lipschitz Continuity of Value Functions.

AMS 90C15 65K05 90C25 65K10 46N10.

1 Introduction and motivation

This work focuses on developing computationally implementable smoothing methods for a family of parametric convex programming problems, noting that all the functions are differentiable but the parametric dependence can be arbitrary.

As one motivating application, the approach provides approximations to (possibly nonconvex) stochastic programs, as long as they exhibit certain structure suitable to our theory. The setting can be illustrated by the following abstract stochastic programming problem formulation:

$$\min_{x\in X} f_0(x) := \mathcal{R}[F(x,\xi(\omega))], \tag{1}$$

where $\mathcal{R}$ is a risk measure [DRS09, Chap. 6], and $F(x,\xi)$ is a real-valued function of the decision variables $x \in X \subset \mathbb{R}^{nx}$. The random vector $\xi(\omega)$ has known probability distribution, with finite support described by scenarios ξ_s and probabilities $p_s \in (0,1)$ for $s = 1,\dots S$.

Pedro Borges
Instituto de Matemática Pura e Aplicada, Rio de Janeiro, RJ, Brazil
E-mail: pborges@impa.br

Claudia Sagastizábal
IMECC/UNICAMP, Campinas, São Paulo, Brazil
E-mail: sagastiz@unicamp.br

Mikhail Solodov
Instituto de Matemática Pura e Aplicada, Rio de Janeiro, RJ, Brazil
E-mail: solodov@impa.br

We start with an example when problem (1) is convex. Consider a two-stage stochastic linear program

$$\begin{cases} \min\ c^\top x + \sum_{s=1}^{S} p_s \mathcal{Q}_s(x) \\ \text{s.t.}\ \ x \in X \end{cases} \quad \text{for}\ \ \mathcal{Q}_s(x) := \begin{cases} \min\ q_s^\top y \\ \text{s.t.}\ \ Wy = h_s - T_s x \\ \qquad y \geq 0\,, \end{cases} \tag{2}$$

where the involved vectors and matrices have suitable dimensions. Suppose the property of relative complete recourse [DRS09, Sect. 2.1.3] is satisfied. Then the format (1) is obtained by taking $\xi = (q_s, h_s, T_s)$, $F(x,\xi_s) = c^\top x + \mathcal{Q}_s(x)$ and $\mathcal{R} = \mathbb{E}$, the expected value function. Since in (2) the first-stage variable appears only in the right-hand side of the feasible set defining the second-stage problems, the corresponding recourse function $\mathcal{Q}_s$ is nonsmooth convex [DRS09, Prop. 2.2]. Hence, so is the associated objective in (1), which is given by

$$f_0(x) = c^\top x + \sum_{s=1}^{S} p_s \mathcal{Q}_s(x)\,. \tag{3}$$

It could be argued that one may get around the nonsmoothness of (2) simply by writing down the deterministic equivalent, a linear programming problem on variables $(x, y_1, \ldots, y_S)$. However, such a rewriting would preclude the possibility of *scenario decomposition* that is present in the nonsmooth formulation. The option to solve separate, easy, second-stage problems (one per scenario s) is very important, and often exploited in real-life applications; [Sag12]. Algorithms based on L-Shaped or bundle methods, [VW69] and [BGLS06, Part II], in particular, generate cuts for the nonsmooth recourse function using the second-stage solutions. The maximum of such cuts is a piecewise affine convex function which by convexity of $\mathcal{Q}_s$ approximates f_0 from below and is used as a proxy in the master program to generate a new first-stage iterate. For such schemes to converge, convexity is fundamental to ensure the generated cuts approximate well the recourse function in regions near the optimum, [OS14; OSL14].

In this work, we shall follow a different path, that is suitable for both convex and certain nonconvex objective functions in (1). The latter setting can occur even when the recourse function $\mathcal{Q}_s$ is convex, if the risk-measure is not convex. An example is [Ahm06, Lem.1], where it is shown that for a stochastic linear program with simple recourse the classical mean-variance criterion yields a piecewise-convex function f_0, which itself is not convex. Risk measures involving the variance are not the only possible source of nonconvexity in (1): the problems considered in [HBT18] have a probability distribution that depends affinely on the first-stage variable. In this case, the function f_0 in (3) is nonsmooth and also nonconvex. Finally, if the second-stage objective function in (2) depends on the first-stage variable, say instead of $q_s^\top y$ we have $q_s(x,y)$, then the recourse function itself can fail to be convex. More instances and examples of similar nature, referred to as programs with linearly bi-parameterized recourse, can be found in [LCPS18].

In order to handle nonsmooth nonconvex objectives, instead of building convex cutting-plane proxies for the recourse function, as in the L-Shaped and bundle methods, we define models that are *smooth and nonconvex*. This is done by adopting a parametric programming point of view, which for (2) amounts to considering the recourse $\mathcal{Q}_s(x)$ as a particular instance of the *value function* of a family of problems that are parameterized by the first-stage variable, x. The proposal replaces each (convex) second-stage problem by a well-behaved strictly or strongly convex (approximating) nonlinear programming problem (NLP), depending on a smoothing parameter $\varepsilon > 0$, and possibly also on a Tikhonov regularization parameter. This NLP unique solution $y_s^\varepsilon(x)$ is a differentiable mapping of x that defines the following smoothed value function:

$$q_s^\top y_s^\varepsilon(x) \geq \mathcal{Q}_s(x)\,, \tag{4}$$

which approximates monotonically the recourse function *from above*. Rather than generating cuts, the master problem minimizes the smoothed objective function

$$c^\top x + \sum_{s=1}^{S} p_s q_s^\top y_s^\varepsilon(x)\,,$$

to define a new first-stage point. An important difference with the L-Shaped family is that now the master program is an NLP. Replacing a piecewise linear master program by a nonlinear version may appear as a handicap at first sight. However, with our scheme not only the solution mappings $y_s^\varepsilon(x)$ are smooth, but they also have *computable derivatives*, related to certain smooth dual mappings, the NLP multipliers computed when solving the approximating second-stage problems. This is a clear algorithmic advantage over the usual cutting-plane models, especially in a nonconvex setting. Additionally, not only (4) holds uniformly for all x

but also, under reasonable conditions, the smoothed recourse function is bounded above by $\mathcal{Q}_s(x)$ plus a term that tends to zero when so does the smoothing parameter; see the relation (11) below.

The smooth approximating solution mappings are defined by suitably combining a Tikhonov regularization with a logarithmic barrier. Regarding related (or somewhat related) works, clearly there are plenty smoothing techniques in the literature. For example, those of [Nes05] and [BT12], which solve convex nonsmooth optimization problems with complexity guarantees. However, complexity analysis is not our subject in this work. The other vastly studied topic concerns generalized and directional derivatives of the optimal value functions; see, e.g., [BS00; RW09] and references therein. Intensive sensitivity analysis of optimization and variational problems via generalized differentiation, including the Lipschitz stability of optimal value/marginal functions, was conducted in [Mor06; Mor18]. Differentiability properties of solution mappings of NLP problems can be traced back to [FM68; Fia83], where the linear independence constraint qualification, strict complementarity, and the second-order sufficient optimality condition are assumed. We must mention here that these works (see also [FI90]) have already considered computing some sensitivity information using approximating penalization schemes. We follow a similar path in the sequel, but with appropriate modifications, among which is adding a Tikhonov regularizing term to the classical interior penalization. Moreover, unlike [FI90], we do not assume satisfaction of strict complementarity or the second-order sufficient condition for the original problem. Accordingly, in our setting the primal solution set need not be a singleton; and can even be unbounded. Instead, we *induce* the second order sufficient condition on certain approximating subproblems, via the specific regularization/penalization scheme employed to compute the approximations. As we shall show, our approach has many interesting theoretical properties, and is also computationally useful; for example, to preserve decomposability of stochastic programs.

As a matter of theory, our regularized penalization scheme provides estimates for the optimal value, as well as locally bounded upper bounds for the subdifferential of the value function. This, in turn, leads to the value function being locally Lipschitz. The latter result recovers, via our computationally-oriented approach and in our case, the locally Lipschitz property established in [GLYZ14] (noting that the setting of [GLYZ14] is much more general). Some other results on the locally Lipschitz behavior of optimal value functions are [MNY09] and [DM15]; but these assume inner semi-continuity of solution mappings (not assumed in this work).

Generally, in this work, we regard smoothing as the ability to generate, in computable ways, single-valued and smooth primal-dual solution mappings, which are "asymptotically correct" in some sense. Therefore, the topic of interest is how the constructed single-valued approximations relate in the limit to the possibly set-valued primal and dual solution mappings, and to the optimal value function.

For various other issues of parametric and sensitivity analysis of optimization and variational problems, see the monographs [FM68; Fia83; BGKKT83; BS00; RW09; Mor06; Mor18], as well as [DGL12]. In [DGL12] the authors analyze certain optimization problems in Banach spaces involving an arbitrary amount of functions that are lower semicontinuous and an abstract constraint given by a closed set. They are mostly concerned with the respective lower and upper continuity aspects of optimal values and solution sets as well as a certain generalized Lipschitz property for the feasible set. To perturb the main optimization problem, they consider a metric on the space of all possible data of the main problem. This amounts to putting a metric on the space of lower semicontinuous functions concatenated with the space of closed sets. They prove that the space of all problem data is complete under their metric. Next, they see feasible set mappings, optimal value functions and solution mappings as mappings on the space of problem data and look at qualitative aspects as well as quantitative relations for these objects. For example, one nontrivial instance for the problems considered in [DGL12] is the master problem of the possibly nonconvex stochastic programming problem considered here. However, we do not deal directly with the underlying nonsmooth problem as opposed to [DGL12]. Instead, we want to show how to build well-behaved smooth approximations to nonsmooth and nonconvex value functions and to understand how these smooth approximations provide some useful information for the value functions. We capitalize on smooth optimization to solve easier approximations for a harder problem, leaving the theoretical issues concentrated solely on how the approximations relate to the original model.

The rest of the paper is organized as follows. In Section 2, we fix notation and blanket assumptions, define the Tikhonov-regularized interior penalty scheme and associated smoothing of the value function, and revise some basic concepts in set-valued analysis. In Section 3, we specialize to our setting several results for the approximate optimal value function, including its parametric differentiability. Some technical bounds are gathered in Section 4. The final theoretical Section 5 shows that the gradients of the approximate value function are locally bounded and, as a by-product of our developments, we recover the result that optimal value functions of qualified convex problems are locally Lipschitz. The developed theory is general, not only applicable to stochastic programs; nevertheless, as two-stage stochastic programs is an important motivation

for us, in Section 6 we go back to the issue of smoothing risk-averse variants of such problems. In the numerical Section 7 the approach is benchmarked on convex instances against a state-of-the-art bundle method software for nonsmooth optimization [Fra02].

2 The setting and main ingredients of the approach

The family of problems in consideration is parameterized by $x \in \mathbb{R}^{nx}$ and has decision variable $y \in \mathbb{R}^{ny}$. The objective function is $f : \mathbb{R}^{nx} \times \mathbb{R}^{ny} \to \mathbb{R}$. Equality constraints are given by means of parametric mappings $A : \mathbb{R}^{nx} \to M(l \times ny)$ and right-hand side maps $b : \mathbb{R}^{nx} \to \mathbb{R}^l$, where $M(l \times nx)$ is the space of $l \times nx$-matrices. The parametric inequality constraints are $g_i : \mathbb{R}^{nx} \times \mathbb{R}^{ny} \to \mathbb{R}$, $i = 1, \ldots, m$. Accordingly, the parametric optimization problem is:

$$\begin{array}{ll} \underset{y}{\text{minimize}} & f(x,y) \\ \text{subject to} & A(x)y = b(x), \\ & g_i(x,y) \leq 0, \quad i = 1, \ldots, m. \end{array} \tag{5}$$

Of course, not all functions need to really have a parametric dependence, and not all types of constraints need to be present. Special cases, like right-hand side and canonical perturbations are included implicitly. In particular, for two-stage linear stochastic programs (2), problem (5) represents the second-stage problems defining the recourse, and only the right-hand side mapping depends on x; specifically, in this case

$$f(x,y) = q_s^\top y,\ A(x) = W,\ b(x) = h_s - T_s x, \text{ and } g(x,y) = -y$$

(so $m = ny$).

2.1 Blanket assumptions and Tikhonov-regularized interior penalty scheme

Throughout we assume that in (5) the following holds for all $x \in \mathbb{R}^{nx}$ (of course, one could instead consider some subset of parameters in $\mathbb{R}^{nx}$):

1. The functions $f(x,\cdot)$ and $g_i(x,\cdot)$, $i = 1, \ldots, m$, are convex.
2. The mappings f, b, A and g_i are at least twice continuously differentiable in both the parameter and the decision variable.
3. The $l \times nx$ matrix $A(x)$ has linearly independent rows.
4. The constraints in (5) satisfy the Slater condition: for every x there exists $\mathring{y}(x)$ such that $A(x)\mathring{y}(x) = b(x)$, $g_i(x,\mathring{y}(x)) < 0$, $i = 1, \ldots, m$.
5. For every x, problem (5) has at least one solution.

Let $S(x)$ denote the (nonempty, possibly unbounded) primal solution set of (5) and let

$$v(x) := f(x, y(x)), \quad \text{for } y(x) \in S(x), \tag{6}$$

be the value function of problem (5).

Recall that [RW09, Theorem 1.17] ensures that the value function (6) is continuous under uniform level-boundedness. While our blanket assumptions above do not imply the latter condition, we can still conclude continuity of the value function via our uniform approximation of the value function via smooth functions, and the condition (14) introduced in the sequel. We do not assume that the Slater points are uniform across parameters. They can change freely with $x \in \mathbb{R}^{nx}$. All the hypotheses about (5) and the associated problem data stated in this section, are not stated again and will be taken in the subsequent sections as granted.

Thus, the main object of our study are smooth parametric convex programming problems, with Slater points and without redundant equality constraints, that have nonempty solution sets for all parameters. The goal is to construct computable (and well-behaved) approximations to primal and dual solution mappings, and to value functions. To that end, define the following *Tikhonov-regularized interior-penalty* (log-barrier) function:

$$\phi(x,y) = -\sum_{i=1}^{m} \ln\{-g_i(x,y)\} + \frac{\mu}{2}\|y\|^2, \tag{7}$$

where $\mu \geq 0$ and $\|\cdot\|$ denotes the Euclidean norm. In our constructions, we use μ fixed, mainly because this turns out to be sufficient for our purposes. In particular, the size of this Tikhonov regularization is controlled by the penalty parameter ε multiplying the full function ϕ, see (8) below. But we could, at the expense of

some extra notation, introduce a separate variable parameter μ_k for the Tikhonov regularization. Moreover, we could also regularize only some variables y_i and not others, depending on the structure of the problem at hand. In particular, the variables that have nonnegativity constraints on them do not need to be regularized, in principle (this would be clear from the subsequent developments). But we shall not go into theoretical analysis of such modifications, as they will cause some technical complications, while the conceptual ideas are clear from our simpler presentation for (7). Note that for the two-stage stochastic linear programs (2), the corresponding penalty function would be $\phi(x,y) = -\sum_{i=1}^{ny} \ln y_i$ (if $\mu = 0$ is taken).

It is worth to point out that the Tikhonov term makes (7) different from the usual log-barrier penalties, but with some similar properties, to be recalled and/or established in the sequel, still holding. At the same time, as we shall explain next, the possibility of adding Tikhonov regularization brings some advantages.

2.2 Regularized approximate value function

For a penalty parameter $\varepsilon > 0$, the Tikhonov-regularized interior penalty approximation of problem (5) is defined by the NLP problem

$$\begin{array}{ll} \underset{y}{\text{minimize}} & f(x,y) - \varepsilon \sum_{i=1}^{m} \ln\{-g_i(x,y)\} + \varepsilon \dfrac{\mu}{2}\|y\|^2 \\ \text{subject to} & A(x)y = b(x) \end{array} \tag{8}$$

(as usual, we use the convention that $\ln t = -\infty$ whenever $t \leq 0$, to drop from (8) the implicit interiority constraints $g_i(x,y) < 0$.)

Our main task is to relate the objects obtained from solving (8) to solutions of (5). To induce the differentiability properties of the interior penalty solutions of (8) we shall assume that either the constraints $y \geq 0$ are present among the inequality constraints in (5), and/or that the regularization parameter $\mu > 0$ is taken in (7). As a result, with our construction, it holds that:

the objective function in (8) is strictly or strongly convex, and its Hessian is positive definite everywhere. (9)

This leads to uniqueness of solutions and eventual differentiability of the solutions mappings. For this reason, when $y \geq 0$ is not present in (5), the Tikhonov term should be added. Otherwise, we do not need to use it, at least if we know that (8) has a solution for $\mu = 0$. The latter is closely related to the solution set of (5) being nonempty and bounded for the given x; see, e.g., [DS99; MZ98] for some results in this direction. But in any case, we can still use the regularization ($\mu > 0$) as well; for example, to make sure that (8) is solvable without any extra assumptions.

The unique solution to the regularized problem (8) defines the estimate of the value function in (6), as follows:

$$v^\varepsilon(x) := f(x, y^\varepsilon(x)), \quad \text{for } y^\varepsilon(x) \text{ solving (8)}. \tag{10}$$

We consider that $v(x) = v^0(x)$, which is justified by the fact that $v^\varepsilon(x) \searrow v(x)$ as $\varepsilon \searrow 0$ (see below for details). We shall also refer to $v^\varepsilon(x)$ as *upper smoothing* (of the value function $v(x)$), which would be justified once it is shown that the mapping $y^\varepsilon(x)$ is differentiable (then so is $v^\varepsilon(x)$). The derivative of $v^\varepsilon(x)$ involves the dual mapping $\lambda^\varepsilon(x)$, the Lagrange multiplier associated to the solution $y^\varepsilon(x)$ of (8). Note that this multiplier is well defined (by the linearity of the constraints in (8)) whenever so is $y^\varepsilon(x)$. In that case, the multiplier is also unique, because $A(x)$ has full row rank, by assumption.

2.3 Basic concepts in set-valued analysis

We now review some notions and relations that will be useful in our development.

Given a set-valued mapping $R : \mathbb{R}^{nx} \to \mathbb{R}^{ny}$, recall that the outer limit of R at $x \in \mathbb{R}^{nx}$ is defined as

$$\limsup_{x \to \bar{x}} R(x) = \{y \in \mathbb{R}^{ny} : \exists x_k \to \bar{x}, \quad y_k \in R(x_k) \quad \text{s.t.} \quad y_k \to y\},$$

and its inner limit by

$$\liminf_{x \to \bar{x}} R(x) := \{y \in \mathbb{R}^{ny} : \quad \forall x_k \to \bar{x} \quad \exists y_k \in R(x_k) \quad \text{s.t.} \quad y_k \to y\}.$$

With a slight abuse of notation, for the value function we shall write $\limsup_{x\to\bar{x}} v(x)$ for $\limsup_{x\to\bar{x}}\{v(x)\}$.

The map R is outer semi-continuous at $\bar{x}\in\mathbb{R}^{nx}$ if $\limsup_{x\to\bar{x}} R(x)\subset R(\bar{x})$. The set-valued map R is said to be inner semi-continuous at $\bar{x}$ if $\liminf_{x\to\bar{x}} R(x)\supset R(\bar{x})$.

A set-valued map R is locally bounded at $\bar{x}\in\mathbb{R}^{nx}$ if there is an open set $V\subset\mathbb{R}^{nx}$ such that $\bar{x}\in V$ and $S(V):=\cup_{x\in V}S(x)$ is bounded.

The regular subdifferential of $v:\mathbb{R}^{nx}\to\mathbb{R}$ at $\bar{x}\in\mathbb{R}^{nx}$ is given by

$$\hat{\partial}v(\bar{x}):=\left\{u\in\mathbb{R}^{nx}:\liminf_{x\to\bar{x}}\frac{v(x)-v(\bar{x})-u^{\top}(x-\bar{x})}{\|x-\bar{x}\|}\geq 0\right\},$$

the limiting subdifferential by

$$\partial v(\bar{x}):=\limsup_{x\to\bar{x}}\hat{\partial}v(x),$$

and the horizon (or singular Mordukhovich) subdifferential by

$$\partial^{\infty}v(\bar{x}):=\left\{u\in\mathbb{R}^{nx}:\exists x_k\to\bar{x},\quad u_k\in\hat{\partial}v(x_k),\quad t_k\searrow 0\quad\text{s.t.}\quad t_k u_k\to u\right\}.$$

Denote by cl D the closure of a set D, and by conv D its convex hull. Then the Clarke subdifferential is given by

$$\partial_C v(x)=\text{conv cl}\,\{\partial v(x)+\partial^{\infty}v(x)\},$$

see [Mor18, Theorem 3.57]. If v is locally Lipschitz, then $\partial_C v(x)=\text{conv}\,\partial v(x)$. To avoid confusion, we mention some alternative terminology widely used in the variational analysis literature: Clarke subdifferential is sometimes called convexified subdifferential (or generalized gradient, in the case of Lipschitz function), regular subdifferential is also known as Fréchet subdifferential, and limiting subdifferential as Mordukhovich subdifferential.

The regular subdifferential and the Clarke subdifferential are convex sets. The Clarke and the limiting subdifferentials are outer semi-continuous multi-functions. When v is convex, it is locally Lipschitz and all these subdifferential notions coincide with the classical subdifferential of convex analysis.

The following proposition characterizes local boundedness of the value-function subdifferential. It is a specialization of [RW09, Theorems 9.13 and 9.2] to our setting (note that in our case the value function is finite-valued).

Proposition 1 (Subdifferential Characterization of Local Lipschitz Continuity)
Let the value function $v:\mathbb{R}^{nx}\to\mathbb{R}$ defined by (5) *be continuous. The following conditions are equivalent:*

1. *The function v is locally Lipschitz at $\bar{x}$.*
2. *The regular subdifferential $\hat{\partial}v$ is locally bounded at $\bar{x}$.*
3. *The limiting subdifferential ∂v is locally bounded at $\bar{x}$.*
4. *The horizon subdifferential $\partial^{\infty}v(\bar{x})$ contains only of the zero vector.*

Proof Theorem 9.13 from [RW09] depends on strict continuity of the value function, given in Definition 9.1, which combined with Theorem 9.2 of [RW09], yields the stated equivalences. Note that strict continuity means local Lipschitz continuity. □

3 The approximate optimal value function and approximating solution mappings differentiability

We now examine how the function (10) approximates the value function (6), and derive formulæ for the derivatives of the associated primal and dual solution mappings $y^{\varepsilon}(x)$ and $\lambda^{\varepsilon}(x)$, respectively.

3.1 Estimates for the optimal value

In this subsection, the analysis concerns a fixed parameter x.

Our penalty approximation (8) of the original problem (5) can be considered to be part of the larger class of interior penalty methods; see [FM68; Wri97]. That said, we are not aware of coupling interior penalties with the Tikhonov regularization, as we do here. Nevertheless, it can be checked directly that certain basic properties hold for this modification as well. In particular, take $0<\varepsilon_2<\varepsilon_1$. As in the classical setting ($\mu=0$, as in [Wri97]), it can be seen that

$$v(x)\leq f(x,y^{\varepsilon_2}(x))\leq f(x,y^{\varepsilon_1}(x))\quad\text{and}\quad\phi(x,y^{\varepsilon_2}(x))\geq\phi(x,y^{\varepsilon_1}(x)).$$

Also, as $\varepsilon \searrow 0$, the accumulation points of $y^\varepsilon(x)$ are solutions of (5), and $v^\varepsilon(x) = f(x, y^\varepsilon(x))$ decreases to $v(x)$, the optimal value of (5). Moreover, when $\mu = 0$, the following uniform estimate for the value function holds:

$$v(x) \le v^\varepsilon(x) \le v(x) + m\varepsilon\,, \quad (\text{ when } \mu = 0)$$

see, e.g., [IS06]. The next proposition generalizes the bound in question to the possibility of using Tikhonov regularization, as in (8).

Proposition 2 (Value function bounds)
For any $\mu \ge 0$ and any $\varepsilon > 0$, if $y^\varepsilon(x)$ exists then it holds that

$$v(x) \le v^\varepsilon(x) \le v(x) + m\varepsilon + \varepsilon\frac{\mu}{2}\min_{y\in S(x)}\|y\|^2. \tag{11}$$

If $\mu > 0$, then $y^\varepsilon(x)$ exists for any $\varepsilon > 0$, and it holds in addition that

$$\frac{\mu}{2}\min_{y\in S(x)}\|y\|^2 + m \ge \frac{\mu}{2}\|y^\varepsilon(x)\|^2. \tag{12}$$

Proof Recall that x is fixed here. Let $\bar\eta_i := -\varepsilon/g_i(x, y^\varepsilon(x)) > 0$ for all $i = 1, \dots, m$. As is easily seen, the KKT optimality conditions for problem (8) characterize $y^\varepsilon(x)$ as a minimizer of

$$L(y) := f(x,y) + \varepsilon\frac{\mu}{2}\|y\|^2 + \sum_{i=1}^m \bar\eta_i g_i(x,y),$$

over the set defined by $A(x)y = b(x)$. Hence, for all $y \in \mathbb{R}^{ny}$ such that $A(x)y = b(x)$, it holds that

$$\begin{aligned} f(x,y) + \varepsilon\frac{\mu}{2}\|y\|^2 + \sum_{i=1}^m \bar\eta_i g_i(x,y) = L(y) &\ge L(y^\varepsilon(x)) \\ &= f(x,y^\varepsilon(x)) + \varepsilon\frac{\mu}{2}\|y^\varepsilon(x)\|^2 + \sum_{i=1}^m \bar\eta_i g_i(x,y^\varepsilon(x)) \\ &= v^\varepsilon(x) + \varepsilon\frac{\mu}{2}\|y^\varepsilon(x)\|^2 - m\varepsilon\,. \end{aligned} \tag{13}$$

Since $\bar\eta_i g_i(x,y) \le 0$ and $f(x,y) = v(x)$ for all $y \in S(x)$, this means that

$$v^\varepsilon(x) \le m\varepsilon + \inf_{y\in S(x)}\left\{f(x,y) + \varepsilon\frac{\mu}{2}\|y\|^2\right\} = v(x) + m\varepsilon + \varepsilon\frac{\mu}{2}\inf_{y\in S(x)}\|y\|^2\,,$$

which is the right-most inequality of (11); while the left-most inequality is obvious.

Similarly, but using also that $v^\varepsilon(x) \ge v(x)$, from (13) we obtain that

$$v(x) + \varepsilon\frac{\mu}{2}\inf_{y\in S(x)}\|y\|^2 = \inf_{y\in S(x)}\left\{f(x,y) + \varepsilon\frac{\mu}{2}\|y\|^2\right\} \ge v(x) + \varepsilon\frac{\mu}{2}\|y^\varepsilon(x)\|^2 - m\varepsilon\,.$$

Now dividing the latter inequality by $\varepsilon > 0$ and re-arranging terms results in (12). □

We do not claim that the estimate (11) is tight, but it will turn out to be sufficient for many purposes. To improve the bound, one would need to estimate how far $y^\varepsilon(x)$ is from the solution in $S(x)$ of minimal norm .

The example $\min\{-y : y \le 1\}$ shows that that the classical bound $v^\varepsilon(x) \le v(x) + m\varepsilon$ is not valid for $\mu > 0$. The claim can be checked explicitly (and is quite clear intuitively), because $\|y^\varepsilon(x)\|^2 < 1 = \min_{y\in S(x)}\|y\|^2$.

Now it is clear that although $\mu > 0$ has the advantage of guaranteeing the existence of $y^\varepsilon(x)$ (and this is without any assumptions), in the parametric context the price to pay is the loss of the uniform approximation of $v(x)$ given by $v^\varepsilon(x)$, because we have to deal with the term $\min_{y\in S(x)}\|y\|^2$ that now appears in the bound.

In the analysis below, for $\bar x \in \mathbb{R}^{nx}$ fixed, we want to know whether $\lim_{\varepsilon\searrow 0, x\to\bar x} v^\varepsilon(x) = v(\bar x)$. Observe that boundedness of the solution set $S(\bar x)$ is not necessarily relevant. For instance, consider $\min\{yx^2 : y \ge 0\}$. We have $S(x) = \{0\}$ for $x > 0$, $S(0) = \{y \in \mathbb{R} : y \ge 0\}$, and $\lim_{\varepsilon\searrow 0, x\to\bar x} v^\varepsilon(x) = v(\bar x)$ holds trivially.

Clearly, one way to ensure that $\lim_{\varepsilon\searrow 0, x\to\bar x} v^\varepsilon(x) = v(\bar x)$ is to guarantee (somehow) that there exists $K > 0$ such that $\min_{y\in S(x)}\|y\|^2 \le K$ for $x \in \mathbb{R}^{nx}$ close to $\bar x$. For instance, if $S(\bar x)$ is locally bounded at $\bar x \in \mathbb{R}^{nx}$, then such $K > 0$ obviously exists. So this is not very restrictive. However, one of the advantages of considering $\mu > 0$ is that we can deal with unbounded solution sets. The constant $K > 0$ in question does not exist if and only if there is a sequence $x_k \to \bar x$ such that $\min_{y\in S(x_k)}\|y\|^2 \to \infty$. This is clearly something rare/pathological,

and can be disregarded in a general approach, like the one we are presenting. Accordingly, where needed, we make the reasonable assumption that

$$\limsup_{x\to\bar{x}}\left\{\min_{y\in S(x)}\|y\|^2\right\}<+\infty. \tag{14}$$

Just note that (14) always holds if the solution sets are locally bounded, which in turn is automatic if the *feasible* sets in (5) are uniformly locally bounded (the latter being quite an acceptable assumption by itself, holding in many cases of interest).

Remark 1 [Consequences of assumption (14)] First, (14) and (11) imply that $v(x)$ is continuous. Under (14), the bound (11) implies $\limsup_{\varepsilon\searrow 0, x\to\bar{x}} y^\varepsilon(x)\subset S(\bar{x})$. However, this does not imply the existence of accumulation points of $y^\varepsilon(x)$. When $\mu>0$ we can use (12) to conclude under (14) that $y^\varepsilon(x)$ remains uniformly bounded for small $\varepsilon>0$ and x close to $\bar{x}\in\mathbb{R}^{nx}$, even if $S(\bar{x})$ is unbounded. The case $\mu=0$ is not as straightforward. If $\mu=0$ we have to assume that $S(x)$ is bounded for all x, so that $y^\varepsilon(x)$ exists. Recalling that convex functions with one nonempty bounded level set are inf-compact and level-bounded, [RW09, Def. 1.8], we can be sure that $y^\varepsilon(x)$ remains bounded for fixed $x\in\mathbb{R}^{nx}$ when we change $\varepsilon>0$. However, to deal with $x\to\bar{x}$ we shall focus on the case when $S(\bar{x})$ is at least locally bounded, if no regularization is used ($\mu=0$). For that, we later refer to the condition

$$\limsup_{\varepsilon\searrow 0, x\to\bar{x}}\|y^\varepsilon(x)\|<+\infty\quad\forall\bar{x}\in\mathbb{R}^{nx}. \tag{15}$$

Satisfaction of (15) is ensured under (14) if $\mu>0$, and under local boundedness of the feasible sets when $\mu=0$. Condition (14) holds, for instance, if the feasible sets are uniformly bounded. An example (due to a referee) for which (14) fails is $\min x^2 y$ s.t. $xy\in[-1,1]$. A more general form of condition (14) is also mentioned in [GLYZ14] as the restricted inf-compactness condition. The difference is that [GLYZ14] allows $S(x)$ to be empty.

3.2 Parametric differentiability

The regularized approximating problem (8) is explicitly set-up to satisfy the associated second-order sufficient optimality condition (by (9), either because of the Tikhonov regularization term with $\mu>0$ or because of the log-barrier penalization of the $y\geq 0$ constraints when they are present). Then, given also the linear independence constraint qualification (by the full rank assumption on the matrices $A(x)$), the differentiability of the mappings $y^\varepsilon(x)$ and $\lambda^\varepsilon(x)$ can be obtained applying to (8) some classical results. We give some details of a direct proof in our case, because the calculations of the derivatives are needed for later developments in any case.

The KKT conditions for (8) give the following parametric system of nonlinear equations (in primal-dual variables):

$$\begin{aligned}&\nabla_y f(x,y^\varepsilon(x))+\varepsilon\nabla_y\phi(x,y^\varepsilon(x))-A(x)^\top\lambda^\varepsilon(x)=0,\\&A(x)y^\varepsilon(x)-b(x)=0.\end{aligned} \tag{16}$$

(Note that we used the constraint in the form of $b(x)-A(x)y=0$ to assign the Lagrange multiplier $\lambda^\varepsilon(x)$ at the solution $y^\varepsilon(x)$ in the first equation above, but then we reversed the sign of the constraint to "the original" in the second equation. This is quite common. Here, we opted for this form for a certain convenience later on.)

Differentiability of the primal-dual solution mappings depends on properties of the Jacobian of (16), which is given by

$$J^\varepsilon(x):=\begin{bmatrix}M^\varepsilon(x) & -A(x)^\top\\ A(x) & 0\end{bmatrix},\ \text{for } M^\varepsilon(x):=\nabla^2_{yy}f(x,y^\varepsilon(x))+\varepsilon\nabla^2_{yy}\phi(x,y^\varepsilon(x)). \tag{17}$$

This is shown below, together with some useful relations to compute the solution mapping derivatives.

Theorem 1 (Smoothness of Solution Mappings)
Let $\varepsilon>0$ be fixed. For all $x\in\mathbb{R}^{nx}$, assume that $y^\varepsilon(x)$ exists (which is automatic if $\mu>0$).
Then the following holds:

(i) The mappings $y^\varepsilon(x)$ and $\lambda^\varepsilon(x)$ are C^1-functions of the parameter $x\in\mathbb{R}^{nx}$.
(ii) For $j=1,\ldots,nx$ the corresponding partial derivatives

$$d^\varepsilon_j(x):=\frac{\partial y^\varepsilon(x)}{\partial x_j}\quad\text{and}\quad\delta^\varepsilon_j(x):=\frac{\partial\lambda^\varepsilon(x)}{\partial x_j} \tag{18}$$

can be computed by solving the linear system

$$J^{\varepsilon}(x)\begin{bmatrix} d_j^{\varepsilon}(x) \\ \delta_j^{\varepsilon}(x) \end{bmatrix} = \begin{bmatrix} \theta_j^{\varepsilon}(x) + \varepsilon\varphi_j^{\varepsilon}(x) \\ \beta_j^{\varepsilon}(x) \end{bmatrix}, \tag{19}$$

where $J^{\varepsilon}(x)$ is given by (17)*, and the right-hand side terms are*

$$\begin{aligned} \theta_j^{\varepsilon}(x) &:= -\frac{\partial \nabla_y f(x,y)}{\partial x_j}\Big|_{y=y^{\varepsilon}(x)} + \frac{\partial A(x)^{\top}}{\partial x_j}\lambda^{\varepsilon}(x), \\ \varphi_j^{\varepsilon}(x) &:= -\frac{\partial \nabla_y \phi(x,y)}{\partial x_j}\Big|_{y=y^{\varepsilon}(x)} \\ \beta_j^{\varepsilon}(x) &:= \frac{\partial b(x)}{\partial x_j} - \frac{\partial A(x)}{\partial x_j}y^{\varepsilon}(x). \end{aligned} \tag{20}$$

Proof To show the first item recall that, by construction, (9) holds, i.e., the matrix $M^{\varepsilon}(x)$ in (17) is positive definite. Take any $(u_1,u_2) \in \ker J^{\varepsilon}(x)$, so that

$$M^{\varepsilon}(x)u_1 - A(x)^{\top}u_2 = 0, \quad A(x)u_1 = 0.$$

Multiplying the first equation above by $u_1^{\top}$ and using $u_1^{\top}A(x)^{\top} = 0$, we conclude that $u_1^{\top}M^{\varepsilon}(x)u_1 = 0$. Positive definiteness of $M^{\varepsilon}(x)$ implies that $u_1 = 0$. Then, by the first equation above, $A(x)^{\top}u_2 = 0$. As $A(x)$ has full row rank, it follows that $u_2 = 0$. Thus $\ker J^{\varepsilon}(x) = \{0\}$, i.e., $J^{\varepsilon}(x)$ is nonsingular.

The conclusions follow from the (second-order) Implicit Function Theorem [Lan93, p. 364]. □

Note that the matrix in the linear systems (19) is the same for all j. This means that only one matrix factorization is required to solve all the linear systems in question.

The next result states that, once the mapping $y^{\varepsilon}(x)$ is smooth, so is the approximating value function $v^{\varepsilon}(x)$, and also gives the expressions for the corresponding derivatives. This justifies the name *upper smoothing* (not to be confused with smoothing in the sense of [Che12]; see Section 5) .

In what follows, for notational simplicity we drop the dependencies of some auxiliary quantities on x and ε, as it is clear from the context.

Corollary 1 (Smoothed value function derivatives)

With the notation and assumptions in Theorem 1, for $i = 1,\ldots,m$ and $j = 1,\ldots,nx$, let

$$\alpha_j := \nabla_y f(x,y^{\varepsilon}(x))^{\top}d_j, \quad \gamma_{ij} := \frac{\nabla_y g_i(x,y^{\varepsilon}(x))^{\top}d_j}{g_i(x,y^{\varepsilon}(x))}.$$

Then it holds that

(i) For each $j = 1,\ldots,nx$,

$$\alpha_j = -\mu\varepsilon y^{\varepsilon}(x)^{\top}d_j + \varepsilon\sum_{i=1}^{m}\gamma_{ij} + \beta_j^{\top}\lambda^{\varepsilon}(x). \tag{21}$$

(ii) The derivatives of the smoothed value function (10) *are given by*

$$\frac{\partial v^{\varepsilon}(x)}{\partial x_j} = \alpha_j + \frac{\partial f(x,y^{\varepsilon}(x))}{\partial x_j},$$

for $j = 1,\ldots,nx$.

Proof Multiplying the transpose of the first identity in (16) by $d_j^{\top}$ gives

$$\alpha_j + \varepsilon\nabla_y\phi(x,y^{\varepsilon}(x))^{\top}d_j - \lambda^{\varepsilon}(x)^{\top}A(x)d_j = 0.$$

For $i = 1,\ldots,m$, define

$$\eta_i := \frac{-\varepsilon}{g_i(x,y^{\varepsilon}(x))}.$$

Taking into account that, by (7),

$$\varepsilon\nabla_y\phi(x,y^{\varepsilon}(x)) = \mu\varepsilon y^{\varepsilon}(x) + \sum_{i=1}^{m}\eta_i\nabla_y g_i(x,y^{\varepsilon}(x)),$$

and $A(x)d_j = \beta_j$ by (19), yields (21).

The second item is just the chain rule, combined with Theorem 1. □

Keeping in mind Proposition 1, formula (11), and the fact that the closure of the smoothed gradients provides an upper bound for the subdifferential of the value function, we would be able to conclude the local Lipschitz continuity of the value function v at a point $\bar{x}$ from boundedness of the gradient of v^ε, by examining the limit of the latter as $x \to \bar{x}$ and $\varepsilon \searrow 0$ (see Section 5 for details). Here, we just point out that in view of Corollary 1, it will suffice to check boundedness of the terms defining the derivatives in item (ii). The right-most term will be dealt with by means of (15), and by smoothness of f and of the regularized solution mapping y^ε. By contrast, bounding the terms α_j is far more involved, and this is the reason for singling out the expression (21) in item (i): the terms therein appear in various inequalities stated in the next section.

4 Technical bounds

The results in this section aim at showing that, although d_j defined in (18) can blow up as $x \to \bar{x}$ and $\varepsilon \searrow 0$, under reasonable conditions the terms α_j defined in Corollary 1 stay bounded (see Theorem 2 below). The strategy we use to do such analysis is not complicated, but the technical details involve many calculations and bounds.

Proposition 3 *With the notation and assumptions in Theorem 1 and Corollary 1, the following relations hold for the matrix $M = M^\varepsilon(x)$ defined in* (17)*, and $\theta_j = \theta_j^\varepsilon(x)$, $\varphi_j = \varphi_j^\varepsilon(x)$, $\delta_j = \delta_j^\varepsilon(x)$ and $\beta_j = \beta_j^\varepsilon(x)$:*

(i) $Md_j = \nabla^2_{yy} f(x, y^\varepsilon(x))d_j + \sum_{i=1}^m \eta_i \nabla^2_{yy} g_i(x, y^\varepsilon(x))d_j - \sum_{i=1}^m \eta_i \gamma_{ij} \nabla_y g_i(x, y^\varepsilon(x)) + \varepsilon\mu d_j\,.$

(ii) $d_j^\top M d_j = d_j^\top \theta_j + \varepsilon d_j^\top \varphi_j + \delta_j^\top \beta_j\,.$

(iii) $d_j^\top M d_j \geq \varepsilon \sum_{i=1}^m \gamma_{ij}^2 + \varepsilon\mu \|d_j\|^2\,.$

Proof Let $\mathbb{I}$ denote the identity matrix of order ny. By the definition of the penalty function in (7),

$$\begin{aligned}\varepsilon \nabla^2_{yy}\phi(x, y^\varepsilon(x)) = &\sum_{i=1}^m \frac{-\varepsilon}{g_i(x, y^\varepsilon(x))} \nabla^2_{yy} g_i(x, y^\varepsilon(x)) \\ &- \sum_{i=1}^m \frac{-\varepsilon}{g_i(x, y^\varepsilon(x))} \nabla_y g_i(x, y^\varepsilon(x)) \frac{\nabla_y g_i(x, y^\varepsilon(x))^\top}{g_i(x, y^\varepsilon(x))} + \varepsilon\mu\mathbb{I}.\end{aligned}$$

The expression in item (i) follows, after multiplying by d_j and recalling the definitions of η_i, γ_{ij}, and of M.

Next, multiplying on the left (19) by the vector $(d_j^\top, \delta_j^\top)$; using the expression in (17) for the Jacobian matrix J, it follows that

$$(d_j^\top, \delta_j^\top) J \begin{bmatrix} d_j \\ \delta_j \end{bmatrix} = \begin{bmatrix} d_j^\top M d_j - d_j^\top A(x)^\top \delta_j \\ \delta_j^\top A(x) d_j \end{bmatrix} = \begin{bmatrix} d_j^\top \theta_j + \varepsilon d_j^\top \varphi_j \\ \delta_j^\top \beta_j \end{bmatrix}.$$

The first line gives item (ii), because $A(x)d_j = \beta_j$.

In the relation for Md_j shown in item (i), the Hessians of f and g_i are positive semidefinite, by convexity of the objective and constraint functions (the implicit constraints $g_i(x,y) < 0$ make $\eta_i > 0$). Accordingly,

$$\begin{aligned} d_j^\top M d_j &\geq -\sum_{i=1}^m \frac{-\varepsilon}{g_i(x, y^\varepsilon(x))} d_j^\top \nabla_y g_i(x, y^\varepsilon(x)) \frac{\nabla_y g_i(x, y^\varepsilon(x))^\top d_j}{g_i(x, y^\varepsilon(x))} + \varepsilon\mu\|d_j\|^2 \\ &= \varepsilon \sum_{i=1}^m \gamma_{ij}^2 + \varepsilon\mu\|d_j\|^2, \end{aligned}$$

which completes the proof. □

The arguments that follow aim at finding upper bounds for the term $d_j^\top M d_j$ in Proposition 3(iii). This is done by bounding from above all the terms in the expression given in Proposition 3(ii). To this aim, our next result states boundedness of $\eta_i = -\varepsilon / g_i(x, y^\varepsilon(x))$, the Lagrange multiplier estimates for inequality constraints, obtained after solving the interior penalty subproblem (8). In the non-parametric case, such results (under appropriate constraint qualifications) are quite classical. Here, we give an extension to the parametric setting of this paper.

But first, we shall need the following property.

Lemma 1 (Continuity of Projections)
Let $Y(x)$ be the feasible set of (5) *for a parameter $x \in \mathbb{R}^{nx}$, and let $\mathring{y}(\bar{x}) \in \mathbb{R}^{ny}$ be a Slater point for the fixed parameter $\bar{x} \in \mathbb{R}^{nx}$.*

It holds that the mapping $P(x)$ of orthogonally projecting the (fixed) point $\mathring{y}(\bar{x})$ onto $Y(x)$ is continuous around $\bar{x} \in \mathbb{R}^{nx}$.

Proof The assertion follows applying [FI90, Theorem 5.1] to the parametric optimization problem

$$\min \|y - \mathring{y}(\bar{x})\|^2 \text{ s.t. } y \in Y(x),$$

where $x \in \mathbb{R}^{nx}$ is the parameter.

Some details. The solution of this problem for the parameter $x = \bar{x}$ is obviously $\mathring{y}(\bar{x})$. By the Slater condition, $g(\bar{x}, \mathring{y}(\bar{x})) < 0$. Hence, the strict complementarity condition and the linear independence of active gradients are automatic for this problem with the parameter $x = \bar{x}$ (the latter because $A(\bar{x})$ has full rank). Finally, the second-order sufficient optimality condition holds by strong convexity of the projecting objective function. Then [FI90, Theorem 5.1] implies that the solution mapping $P(x)$ of the problem in question is smooth around $\bar{x}$. □

Remark 2 If the point $\mathring{y}(\bar{x})$ in Lemma 1 were to be changed to an arbitrary (but fixed) point, we could still use the inequality (11), written for the projection problem, to conclude that $P(x)$ is continuous if the projection $P(x)$ is locally bounded. This is possible because inequality (11) shows that there is a sequence of smooth functions converging locally uniformly to $P(x)$.

Lemma 2 (Local Boundedness of Multiplier Estimates η_i)
Assume that $y^{\varepsilon}(x)$ exists for all $x \in \mathbb{R}^{nx}$ (which is automatic if $\mu > 0$), and that (15) *holds at $\bar{x} \in \mathbb{R}^{nx}$. Then for all $i = 1, \ldots, m$,*

$$0 \leq \limsup_{\varepsilon \searrow 0, x \to \bar{x}} \eta_i < +\infty. \tag{22}$$

Proof To show (22) suppose, for contradiction purposes, that there exist $\varepsilon_k \searrow 0$ and $x_k \to \bar{x}$ such that for some $i = 1, \ldots, m$ it holds that $\{-\varepsilon_k / g_i(x_k, y^{\varepsilon_k}(x_k))\} \to +\infty$. Taking subsequences of $\{\varepsilon_k\}$ and $\{x_k\}$ we can get a partition of $\{1, \ldots, m\} = I_0 \cup I_\infty$ where for all $i \in I_0$ the sequences $\{-\varepsilon_k / g_i(x_k, y^{\varepsilon_k}(x_k))\}$ remain bounded, while

$$-\varepsilon_k / g_i(x_k, y^{\varepsilon_k}(x_k)) \to +\infty \text{ for } i \in I_\infty. \tag{23}$$

Denote by $Y(x)$ the feasible set of (5) for a parameter $x \in \mathbb{R}^{nx}$. Let $\mathring{y}(\bar{x}) \in \mathbb{R}^{ny}$ be a Slater point for the parameter $\bar{x} \in \mathbb{R}^{nx}$ (which exists by the blanket assumptions). Define $y_k = \mathrm{P}_{Y(x_k)}(\mathring{y}(\bar{x}))$ to be the projection of $\mathring{y}(\bar{x})$ onto $Y(x_k)$. By Lemma 1, we have that $y_k \to \mathring{y}(\bar{x})$. Also by continuity, there exists some $\Gamma > 0$ such that, for all k large enough,

$$g_i(x_k, y_k) \leq -\frac{\Gamma}{2} < 0 \quad \text{for all } i \in I_0 \cup I_\infty, \tag{24}$$

because $g_i(\bar{x}, \mathring{y}(\bar{x})) \leq -\Gamma < 0$ for all $i \in I_0 \cup I_\infty$.

Define

$$u_k := y_k - y^{\varepsilon_k}(x_k),$$

and note that

$$A(x_k) u_k = 0.$$

Take $i \in I_\infty$. By (23), we have that $g_i(x_k, y^{\varepsilon_k}(x_k)) \to 0$. By convexity,

$$g_i(x_k, y_k) \geq g_i(x_k, y^{\varepsilon_k}(x_k)) + [\nabla_y g_i(x_k, y^{\varepsilon_k}(x_k))]^\top u_k.$$

Using $g_i(x_k, y^{\varepsilon_k}(x_k)) \to 0$ and (24), we can assume that for all k large enough it holds that

$$\nabla_y g_i(x_k, y^{\varepsilon_k}(x_k))^\top u_k \leq -\frac{\Gamma}{4} < 0 \text{ for all } i \in I_\infty. \tag{25}$$

Multiplying on the left by $u_k^\top$ the KKT condition (16) written with (ε_k, x_k), we see that

$$u_k^\top \nabla_y f(x_k, y^{\varepsilon_k}(x_k)) + \varepsilon_k \nabla_y \phi(x_k, y^{\varepsilon_k}(x_k))^\top u_k = u_k{}^\top [A(x_k)^\top \lambda^{\varepsilon_k}(x_k)] = 0,$$

because $A(x_k) u_k = 0$. Since by (7),

$$\varepsilon_k \nabla_y \phi(x_k, y^{\varepsilon_k}(x_k)) = \sum_{i \in I_0 \cup I_\infty} \eta_i \nabla_y g_i(x_k, y^{\varepsilon_k}(x_k)) + \mu \varepsilon_k y^{\varepsilon_k}(x_k),$$

it follows that

$$u_k^\top \nabla_y f(x_k, y^{\varepsilon_k}(x_k)) - \varepsilon_k \sum_{i \in I_0 \cup I_\infty} \frac{\nabla_y g_i(x_k, y^{\varepsilon_k}(x_k))^\top u_k}{g_i(x_k, y^{\varepsilon_k}(x_k))} + \mu \varepsilon_k u_k^\top y^{\varepsilon_k}(x_k) = 0\,.$$

Hence,

$$\begin{aligned} &-\varepsilon_k \sum_{i \in I_\infty} \frac{\nabla_y g_i(x_k, y^{\varepsilon_k}(x_k))^\top u_k}{g_i(x_k, y^{\varepsilon_k}(x_k))} \\ &= -u_k^\top \nabla_y f(x_k, y^{\varepsilon_k}(x_k)) + \varepsilon_k \sum_{i \in I_0} \frac{\nabla_y g_i(x_k, y^{\varepsilon_k}(x_k))^\top u_k}{g_i(x_k, y^{\varepsilon_k}(x_k))} - \mu \varepsilon_k u_k^\top y^{\varepsilon_k}(x_k)\,. \end{aligned}$$

The left-hand side in the equality above tends to $-\infty$ as $k \to \infty$, by (23) and (25). The sequence $\{y^{\varepsilon_k}(x_k)\}$ is bounded because of (15). Then, $\{u_k\}$ is also bounded, as well as all the terms in the right-hand side of the equality above. Thus, we have a contradiction. This proves (22). □

We then obtain the following.

Corollary 2 (Local Boundedness of the Smoothed Dual Solution Mapping)
Under the assumptions of Lemma 2, for any $\bar{x} \in \mathbb{R}^{nx}$ there exist $\rho > 0$ and $C > 0$ such that, for all $\varepsilon \in (0, \rho)$ and $x \in B(\bar{x}, \rho)$, it holds that

(i) $\|Md_j\| \le C\|d_j\| + C\sum_{i=1}^{m} |\gamma_{ij}|$.

(ii) *The set $\{\lambda^\varepsilon(x) : x \in B(\bar{x}, \rho)\,, \varepsilon \in (0, \rho]\}$ is bounded.*

Proof By Proposition 3(i),

$$\begin{aligned} \|Md_j\| \le &\|\nabla^2_{yy} f(x, y^\varepsilon(x))\| \|d_j\| + \sum_{i=1}^{m} \eta_i \|\nabla^2_{yy} g_i(x, y^\varepsilon(x))\| \|d_j\| + \varepsilon\mu\|d_j\| \\ &+ \sum_{i=1}^{m} \eta_i \left|\gamma_{ij}\right| \|\nabla_y g_i(x, y^\varepsilon(x))\|\,. \end{aligned}$$

Item (i) follows, by (15) and (22).

As the matrices $A(x)$ have full rank, from the KKT conditions (16) we obtain, in a standard way, that

$$\lambda^\varepsilon(x) = [A(x)A^\top(x)]^{-1} A(x) \left\{ \nabla_y f(x, y^\varepsilon(x)) + \varepsilon\mu y^\varepsilon(x) + \sum_{i=1}^{m} \eta_i \nabla_y g_i(x, y^\varepsilon(x)) \right\}.$$

Item (ii) follows, again by (15) and (22). □

Keeping Proposition 3(ii) in mind, we next estimate the behavior of the right-hand side terms in the linear system (19).

Proposition 4 *Under the assumptions of Lemma 2, for all $j = 1, \ldots, nx$ and $\bar{x} \in \mathbb{R}^{nx}$, the quantities $\theta_j = \theta_j^\varepsilon(x)$ and $\varphi_j = \varphi_j^\varepsilon(x)$, defined in* (20)*, satisfy the following relations:*

(i) $\limsup_{\varepsilon \searrow 0, x \to \bar{x}} \varepsilon^2 \|\varphi_j\| < +\infty$ *and* $\limsup_{\varepsilon \searrow 0, x \to \bar{x}} \|\theta_j\| < +\infty$.

(ii) $\varepsilon^2 |d_j^\top \varphi_j| \le \varepsilon K(\|d_j\| + \sum_{i=1}^{m} \left|\gamma_{ij}\right|)$ *for $\varepsilon \in (0, \delta)$, $x \in B(\bar{x}, \delta)$ and some constant $K = K(\delta, \bar{x}) > 0$.*

Proof Recalling the definition of φ_j, we obtain that

$$\begin{aligned} \varepsilon^2 \varphi_j &= -\sum_{i=1}^{m} \frac{\partial \left(\varepsilon \eta_i \nabla_y g_i(x, y) \right)}{\partial x_j} \Big|_{y = y^\varepsilon(x)} \\ &= -\varepsilon \sum_{i=1}^{m} \frac{\partial \eta_i}{\partial x_j} \nabla_y g_i(x, y^\varepsilon(x)) - \varepsilon \sum_{i=1}^{m} \left(\eta_i \frac{\partial \nabla_y g_i(x, y^\varepsilon(x))}{\partial x_j} \right). \end{aligned}$$

We now bound the right-hand side terms, as follows. First notice that by (15) and (22), for some $K_1 > 0$ (depending on $\bar{x}$) it holds that for all x close enough to $\bar{x}$ and all ε close enough to zero,

$$\left\| \varepsilon \sum_{i=1}^{m} \left(\eta_i \frac{\partial \nabla_y g_i(x, y^\varepsilon(x))}{\partial x_j} \right) \right\| \le K_1\,. \tag{26}$$

Regarding the terms in the first summation, recalling the definition of η_i,

$$\begin{aligned}\varepsilon\frac{\partial\eta_i}{\partial x_j} &= -\varepsilon^2\frac{\partial\left(1/g_i(x,y)\right)}{\partial x_j}\Big|_{y=y^\varepsilon(x)}\\ &= \frac{\varepsilon^2}{(g_i(x,y^\varepsilon(x)))^2}\frac{\partial g_i(x,y)}{\partial x_j}\Big|_{y=y^\varepsilon(x)}\\ &= (\eta_i)^2\frac{\partial g_i(x,y)}{\partial x_j}\Big|_{y=y^\varepsilon(x)}.\end{aligned}$$

Using once more (22), together with smoothness of g_i and (15), we conclude that the term above is bounded. Therefore, there exists some constant $K_2>0$ such that

$$\left\|\varepsilon\sum_{i=1}^m\frac{\partial\eta_i}{\partial x_j}\nabla_y g_i(x,y^\varepsilon(x))\right\|\le K_2\,. \tag{27}$$

Combining (26) with (27) gives the first assertion in item (i).

The second assertion in item (i) follows using the smoothness assumptions on f and A, (15), and item (ii) of Corollary 2.

Item (ii) follows multiplying the expression above for $\varepsilon^2\varphi_j$ by d_j, and re-examining the terms involved. □

The final estimate of this section is the following.

Proposition 5 *Under the assumptions of Lemma 2, for all $j=1,\dots,nx$ and $\bar{x}\in\mathbb{R}^{nx}$ there exist $\rho>0$ and a constant $L>0$ such that, for all $\varepsilon\in(0,\rho)$ and $x\in B(\bar{x},\rho)$,*

$$\varepsilon^2\mu\|d_j\|^2+\varepsilon^2\sum_{i=1}^m\gamma_{ij}^2\le\varepsilon L\|d_j\|+\varepsilon L\sum_{i=1}^m\left|\gamma_{ij}\right|+L. \tag{28}$$

Proof Throughout we consider $\varepsilon>0$ sufficiently small and x close enough to $\bar{x}\in\mathbb{R}^{nx}$. We also drop the dependencies on ε and x, as it is clear from the context. For example, in what follows $M:=M^\varepsilon(x)$, $A:=A(x)$, as well as $d_j:=d_j^\varepsilon(x)$, $\delta_j:=\delta_j^\varepsilon(x)$, etc.

By items (ii) and (iii) in Proposition 3, we have that

$$\varepsilon^2\mu\|d_j\|^2+\varepsilon^2\sum_{i=1}^m\gamma_{ij}^2\le\varepsilon d_j^\top\theta_j+\varepsilon^2 d_j^\top\varphi_j+\varepsilon\delta_j^\top\beta_j\,. \tag{29}$$

To establish (28), we proceed to bound the terms in the right-hand side of (29).

By items (i) and (ii) in Proposition 4, for some constant $L_1>0$,

$$\varepsilon d_j^\top\theta_j+\varepsilon^2 d_j^\top\varphi_j\le\varepsilon\|\theta_j\|\|d_j\|+\varepsilon^2|d_j^\top\varphi_j|\le\varepsilon L_1(\|d_j\|+\sum_{i=1}^m\left|\gamma_{ij}\right|)\,. \tag{30}$$

To bound the last term in the right-hand side of (29), we first show that, for some constant $L_2>0$,

$$\varepsilon\|\delta_j\|\le\varepsilon L_2\|d_j\|+\varepsilon L_2\sum_{i=1}^m\left|\gamma_{ij}\right|+L_2\,. \tag{31}$$

By the first equation in (19), $Md_j-A^\top\delta_j=\theta_j+\varepsilon\varphi_j$. Multiplying this equation by A, as the matrix $AA^\top$ is non-singular, we obtain that

$$\delta_j=(AA^\top)^{-1}A(Md_j-\theta_j-\varepsilon\varphi_j).$$

Since the matrices $(AA^\top)^{-1}A$ (which depend on x) are bounded for all x close to $\bar{x}$, for some $L_2>0$

$$\begin{aligned}\varepsilon\|\delta_j\| &\le L_2(\|\varepsilon Md_j\|+\varepsilon\|\theta_j\|+\varepsilon^2\|\varphi_j\|\\ &\le L_2\varepsilon\|Md_j\|+L_3,\end{aligned} \tag{32}$$

where the second inequality follows from Proposition 4(i), taking $L_3>0$ large enough.

By item (i) in Corollary 2, for some constant $C>0$,

$$\|Md_j\|\le C\|d_j\|+C\sum_{i=1}^m\left|\gamma_{ij}\right|.$$

Combining the latter relation with (32) and taking $L_2 > 0$ large enough, gives (31).

By the definition of β_j and (15), $\|\beta_j\|$ stays bounded as $\varepsilon \searrow 0$ and $x \to \bar{x}$. By (31), it then holds that, for some $L_4 > 0$,

$$\varepsilon \delta_j^\top \beta_j \leq \varepsilon \|\delta_j\| \|\beta_j\| \leq \varepsilon L_4 \|d_j\| + \varepsilon L_4 \sum_{i=1}^{m} \left|\gamma_{ij}\right| + L_4 .$$

Combining the latter relation with (30), the assertion (28) follows from (29). □

5 Boundedness of the smoothing gradients and Lipschitz-continuity of the value function

We shall now discuss some consequences of our analysis above, including boundedness of the derivatives of the proposed smoothing, as well as some issues related to gradient consistency [Che12; BHK13; BH17; BH13], and Lipschitz-continuity of the value function [MNY09; DM15; GLYZ14].

We are now in position to combine the various inequalities in Section 4 to bound the upper smoothing derivatives given in Corollary 1.

Theorem 2 (Local Uniform Boundedness of Smoothed Gradient)
Assume that the smoothing is built with a fixed $\mu > 0$ and that (14) *holds at $\bar{x} \in \mathbb{R}^{nx}$. Then there exist $\rho > 0$ and $L > 0$ such that*

$$\|\nabla v^\varepsilon(x)\| \leq L \quad \textit{for all } \varepsilon \in (0,\rho) \textit{ and } x \in B(\bar{x},\rho).$$

Proof Take any $j \in \{1,\ldots,nx\}$. Recalling Corollary 1, we have that

$$\frac{\partial v^\varepsilon(x)}{\partial x_j} = -\mu \varepsilon y^\varepsilon(x)^\top d_j + \varepsilon \sum_{i=1}^{m} \gamma_{ij} + \beta_j^\top \lambda^\varepsilon(x) + \frac{\partial f(x, y^\varepsilon(x))}{\partial x_j} . \tag{33}$$

The last term in the right-hand side of (33) is locally bounded by the assumption (15), and the smoothness properties of f and of y^ε (the latter established in Theorem 1). As already used before, β_j is also bounded, by the same reasons. The mapping λ^ε is locally bounded, as established in Corollary 2(ii). Hence, the last two terms in the right-hand side of (33) are bounded. It remains to analyze the first two terms.

Suppose that the term $\varepsilon\|d_j\|$ is unbounded as $\varepsilon \searrow 0$ and $x \to \bar{x}$. By (28), it holds that

$$\varepsilon^2 \mu \|d_j\|^2 \leq \varepsilon L \|d_j\| + \varepsilon L \sum_{i=1}^{m} \left|\gamma_{ij}\right| + L .$$

As $\mu > 0$, this inequality implies that if $\varepsilon\|d_j\|$ is unbounded, then the term $\varepsilon \sum_{i=1}^{m} \left|\gamma_{ij}\right|$ must be unbounded (otherwise the inequality in question yields a contradiction). But both $\varepsilon\|d_j\|$ and $\varepsilon \sum_{i=1}^{m} \left|\gamma_{ij}\right|$ being unbounded clearly contradicts (28), recalling again that $\mu > 0$. We conclude that $\varepsilon\|d_j\|$ is bounded. Then (28) implies that so is $\varepsilon \sum_{i=1}^{m} \left|\gamma_{ij}\right|$.

The proof is completed, because we showed that all the terms in the right-hand side of (33) are bounded. □

Note, in passing, that the analysis above shows that the following bound on the possible blow-up rate of the derivatives of y^ε holds:

$$\limsup_{\varepsilon \searrow 0, x \to \bar{x}} \varepsilon \|\nabla y^\varepsilon(x)\| < +\infty.$$

We next make some comments on other notions appearing in the literature on smoothing, and in particular on the property known as *gradient consistency*. Gradient consistency was introduced in [CQS98] as Jacobian consistency, and further studied in [BHK13] and [Che12]; see also [QSZ00; RX05].

In the following definitions, a continuous function, possibly nonsmooth, $v : \mathbb{R}^{nx} \to \mathbb{R}$ is given, as well as a differentiable function $\sigma : (0,\infty) \times \mathbb{R}^{nx} \to \mathbb{R}$.

– The smooth function σ is said to be a *smoothing of v in the sense of [Che12]* if

$$\lim_{\varepsilon \searrow 0, x \to \bar{x}} \sigma(\varepsilon, x) = v(\bar{x}). \tag{34}$$

– When v is locally Lipschitz, the property of *gradient consistency* between σ and v holds, see [Che12], whenever

$$\operatorname{conv} \left\{ \limsup_{\varepsilon \searrow 0, x \to \bar{x}} \nabla_x \sigma(\varepsilon, x) \right\} \subset \partial_C v(\bar{x}). \tag{35}$$

Note that it is an easy consequence of [BHK13, Lemma 3.1], that for every smoothing function σ of v (smoothing in the sense of (34)), the "almost" converse of the inclusion (35) always holds, i.e.,

$$\operatorname{conv}\left\{\limsup_{\varepsilon\searrow 0, x\to\bar{x}} \nabla_x \sigma(\varepsilon,x)\right\} \supset \partial v(\bar{x}). \tag{36}$$

In our context v is the optimal value function of problem (5) while, for the given regularization/penalization parameter $\varepsilon > 0$ appearing in (8), we have $\sigma(\varepsilon,x) = v^\varepsilon(x) = f(y^\varepsilon(x),x)$, with $y^\varepsilon(x)$ being the solution of problem (8).

When gradient consistency was defined in [Che12], the motivating smoothing functions considered there were explicit, and so local boundedness of the gradients was something granted, in a sense. In our case, the situation is different (as our smoothing function is implicit), and indeed we had to prove that its gradients remain bounded. In general, boundedness/unboundedness is relevant, because of the formula for the Clarke subdifferential that involves the horizon subdifferential (see Section 2.3). This information is not present ("missing") in (35) and (36), as those conditions are intended for bounded sequences of gradients. In this paper (as a side issue, not our principal concern) we prove that the horizon subdifferential of v and horizon closure of the smoothed gradients ($\limsup^\infty$) are equal, which is part of what is needed in the general gradient consistency theory, beyond the current definition (35) for locally Lipschitz functions (where the smoothing function has locally bounded gradients). For instance, consider the problem $v(x) = \min_y xy$ s.t. $y \geq 0$ and the smoothing of the value function with $\mu > 0$. There are unbounded smoothed gradients around $x = 0$. Clearly, our assumptions do not hold for this v because $S(x) = \emptyset$ if $x < 0$.

In the statements below, we make use of the condition (14), whose consequences were discussed in Remark 1.

Lemma 3 (Gradient Consistency)

The following holds true:

(i) If $\mu = 0$, or if $\mu > 0$ and (14) *holds, then the function v^ε is a smoothing for v in the sense of [Che12] (i.e.,* (34) *holds for $\sigma(\varepsilon,x) = v^\varepsilon(x)$).*

(ii) If $\mu > 0$ and (14) *holds, then $w^\varepsilon(x) := v^\varepsilon(x) + \varepsilon\phi(x, y^\varepsilon(x))$ is also a smoothing in the sense of [Che12], and it has locally bounded gradients.*

(iii) If problem (5) *has parameters only on the map b and b is affine, (i.e.,* (5) *has only right-hand side linear perturbations), then v and w^ε are convex.*

(iv) Under the assumptions of Theorem 2, if v is convex and w^ε is convex for $\varepsilon > 0$ small enough, then w^ε is gradient consistent with v (i.e., (35) *holds for $\sigma(\varepsilon,x) = w^\varepsilon(x)$).*

Proof When $\mu = 0$, or (14) holds for $\mu > 0$, the relation (11) in Proposition 2 and the continuity of v imply that

$$\lim_{\varepsilon\searrow 0, x\to\bar{x}} v^\varepsilon(x) = \lim_{x\to\bar{x}} v(x) = v(\bar{x}).$$

Item (i) follows.

Let us now prove item (ii). In view of item (i), to show that w^ε is a smoothing of v we have to verify that $\varepsilon\phi(x', y^\varepsilon(x')) \to 0$ when $\varepsilon \searrow 0$ and $x' \to x$. Recall that $\varepsilon\mu\|y^\varepsilon(x')\| \to 0$ when $\varepsilon \searrow 0$ and $x' \to x$, due to (14) and (12). Next, using Lemma 2 and (14), we can conclude that $\varepsilon \ln\{-g_i(x', y^\varepsilon(x'))\} \to 0$ since there is $C > 0$ such that $\varepsilon \leq -Cg_i(x', y^\varepsilon(x'))$ and $y^\varepsilon(x')$ is bounded. We conclude that $\varepsilon\phi(x', y^\varepsilon(x')) \to 0$, and thus w^ε is a smoothing of v.

Computing the gradient of w^ε via the chain rule we see that it is locally bounded, because εd_j and $\varepsilon\gamma_{ij}$ are bounded (see the proof of Theorem 2), and the other terms can be bounded using (14), (12) and Lemma 2.

We proceed to item (iii). Consider the problem $\tilde{v}(x) = \min_y \tilde{f}(y)$ s.t. $Ay = b(x)$, where $\tilde{f}(y)$ is a convex extended-valued function, and the problem has solutions for all x. For v we shall have $\tilde{f} = f + I_D$, where I_D is the indicator function of the set $D = \{x : g(x) \leq 0\}$, and for w^ε we have $\tilde{f} = f + \varepsilon\phi$. Denote by $\tilde{y}(x)$ any solution for a fixed x. Taking any x_1, x_2 and $t \in (0,1)$, the point $t\tilde{y}(x_1) + (1-t)\tilde{y}(x_2)$ is feasible for the problem at parameter $x = tx_1 + (1-t)x_2$. It follows by the convexity of $\tilde{f}$ that $\tilde{v}(tx_1 + (1-t)x_2) \leq \tilde{f}(t\tilde{y}(x_1) + (1-t)\tilde{y}(x_2)) \leq t\tilde{v}(x_1) + (1-t)\tilde{v}(x_2)$. This shows convexity of $\tilde{v}$, i.e., of v and w^ε.

To establish item (iv), fix $\bar{x} \in \mathbb{R}^{nx}$. By the convexity of w^ε, for any x and x' it holds that

$$w^\varepsilon(x') \geq w^\varepsilon(x) + \nabla w^\varepsilon(x)^\top (x' - x).$$

Note that $\lim_{\varepsilon \searrow 0} w^{\varepsilon}(x') = v(x')$ and $\lim_{\varepsilon \searrow 0, x \to \bar{x}} w^{\varepsilon}(x) = v(\bar{x})$. Then, passing onto the limits $\varepsilon \searrow 0$ and $x \to \bar{x}$ in the inequality above, for any $u \in \limsup_{\varepsilon \searrow 0, x \to \bar{x}} \nabla w^{\varepsilon}(x)$ we see that it must hold that

$$v(x') \geq v(\bar{x}) + u^{\top}(x' - \bar{x}).$$

(Note that such u exists, by item (ii).) This shows that u is a subgradient of the convex function v at $\bar{x}$, implying gradient consistency. □

Remark 3 Note that our smoothing can also be seen in the context of Attouch's Theorem [Att77]. Then, Lemma 3 could be viewed as an "implementation" of Attouch's Theorem, in the sense that our approximating functions are computable.

We finish by showing that the optimal value function v is locally Lipschitz under our assumptions. We note that a more general result is available in [GLYZ14]. However, here we obtain the locally Lipschitz property of v as a simple by-product of our algorithmic smoothing approach.

Theorem 3 (Local Lipschitz Continuity of the Value Function)
In addition to the blanket assumptions stated in Section 2, assume that condition (14) *holds for* $\bar{x} \in \mathbb{R}^{nx}$. *Then the optimal value function* v *is locally Lipschitz continuous in the neighborhood of* $\bar{x}$.

Proof Take any $\mu > 0$ and consider (8), which in this case always has solution $y^{\varepsilon}(x)$, for every $\varepsilon > 0$. As (14) is assumed, by Lemma 3 we know that the corresponding v^{ε} is a smoothing of v. Hence, by (36), it holds that

$$\partial v(\bar{x}) \subset \operatorname{conv} \left\{ \limsup_{\varepsilon \searrow 0, x \to \bar{x}} \nabla v^{\varepsilon}(x) \right\}. \tag{37}$$

Next, as explained in Remark 1, when $\mu > 0$, condition (14) implies (15) (because of (12)). Then, by Theorem 2, $\nabla v^{\varepsilon}(x)$ is locally bounded. Hence, by (37), so is $\partial v(\bar{x})$.

The conclusion now follows from Proposition 1. □

Note that in Theorem 3, taking $\mu > 0$ is useful for providing a locally bounded upper bound for $\partial v(x)$, and the resulting theoretical argument. This is not related to choosing μ in any computational implementation of the smoothing approach.

6 Smoothing risk-averse two-stage stochastic programs

We now explain how to cast in our setting a two-stage convex stochastic program, to be considered in our computational experiments in Section 7.

Given a risk-aversion parameter $\kappa \in [0,1]$ and a confidence level $\alpha \in (0,1)$, we combine expected value with average-value-at-risk functionals to define

$$\mathcal{R}[Z] := \kappa \mathbb{E}[Z] + (1-\kappa)\mathtt{AVaR}_{\alpha}(Z),$$

for a random variable Z representing a loss ($\kappa = 1$ is the risk-neutral variant). Letting $c(x)$ and $q_s(y)$ denote convex first and second-stage objective functions, the risk-averse two-stage stochastic program of interest is

$$\begin{cases} \min\ c(x) + \mathcal{R}\left[q_1(y_1), \ldots, q_S(y_S)\right] \\ \text{s.t.}\ \ x \in X \\ \qquad \text{and, for } s = 1, \ldots, S \\ \qquad y_s \geq 0, T_s x + W y_s = h_s, \end{cases} \tag{38}$$

where we assume once more that the recourse is relatively complete, so that the second-stage problems have nonempty feasible sets. Using the expression

$$\mathtt{AVaR}_{\alpha}[Z] := \min_{x_u \in \mathbb{R}} \left\{ x_u + \frac{1}{1-\alpha} \mathbb{E}\left[\max(Z - x_u, 0)\right] \right\}$$

from [RU02], we obtain the following risk-averse version of the two-level problem (2):

$$\begin{cases} \min\ c(x) + (1-\kappa)x_u + \displaystyle\sum_{s=1}^{S} p_s \mathcal{Q}_s(x, x_u) \\ \text{s.t.}\ \ x \in X, x_u \in \mathbb{R}, \end{cases} \quad \text{for } \mathcal{Q}_s(x, x_u) := \begin{cases} \min\ \kappa q_s(y) + \dfrac{1-\kappa}{1-\alpha} z \\ \text{s.t.}\ \ W y = h_s - T_s x \\ \qquad q_s(y) - z \leq x_u \\ \qquad y \geq 0, z \geq 0 \end{cases} \tag{39}$$

By construction, the second-stage objective function and constraints in (39) are convex on (y,z), while the recourse function is finite-valued, nonsmooth and convex on (x,x_u). Furthermore, the optimal multipliers of the constraints involving (x,x_u), say η_x, η_{x_u} with $\eta_{x_u} \geq 0$, provide the subgradient $(T_s^\top \eta_x, -\eta_{x_u})^\top$.

For $s = 1,\ldots,S$, the smoothed second-stage solutions, denoted $\left(y^\varepsilon(x,u), z^\varepsilon(x,u)\right)$, are computed by solving the smoothed second-stage problems

$$\begin{cases} \min\ \kappa q_s(y) + \dfrac{1-\kappa}{1-\alpha} z + \varepsilon \phi_s(x_u,y,z) \\ \text{s.t.}\ \ Wy = h_s - T_s x, \end{cases} \quad \text{for } \phi_s(x_u,y,z) := -\sum_{i=1}^{ny} \ln(y_i) - \ln(z) - \ln(z - q_s(y) + x_u). \tag{40}$$

The approximate first-stage problem is

$$\begin{cases} \min\ c(x) + (1-\kappa)x_u + \displaystyle\sum_{s=1}^{S} p_s \left(\kappa q_s(y^\varepsilon(x,x_u)) + \frac{1-\kappa}{1-\alpha} z^\varepsilon(x,x_u) \right) \\ \text{s.t.}\ \ x \in X, x_u \in \mathbb{R}, \end{cases} \tag{41}$$

which, by the definition of $\texttt{AVaR}_\alpha$, is not necessarily the same as the objective of the problem below:

$$\begin{cases} \min\ c(x) + \mathcal{R}\left[q_1(y_1^\varepsilon(x,x_u)), \ldots, q_S(y_S^\varepsilon(x,x_u)) \right] \\ \text{s.t.}\ \ x \in X. \end{cases}$$

Corollary 3 (Specializing the Results to Two-Stage Risk-Averse Stochastic Linear Programs)
Consider the particular instance of the abstract stochastic problem (1) *given by* (39) *and its smooth approximation* (41). *Suppose that the matrix W has linearly independent rows. Assume also that for all $x \in X$ the recourse problems, without risk measures, satisfy the Slater condition and have nonempty solution sets. Then the following holds when building the smoothing with $\mu = 0$ as in* (40)*:*

(i) For $s = 1,\ldots,S$,

$$Q_s(x,x_u) \leq \kappa q_s(y^\varepsilon(x,x_u)) + \frac{1-\kappa}{1-\alpha} z^\varepsilon(x,x_u) \leq Q_s(x,x_u) + \varepsilon C_s$$

for an explicit and known constant $C_s > 0$.

(ii) The objective function of (41) *decreases monotonically and uniformly to the objective function of* (39) *as $\varepsilon \searrow 0$.*

(iii) If $\overline{x}^\varepsilon$ is a global solution to (41) *then $\overline{x}^\varepsilon$ is an approximate global solution to* (39) *with explicit and known quality of approximation.*

Proof To prove item (iii), look at item (i). Start multiplying (i) by p_s, and then summing across the scenarios. After that, add the first-stage cost $c(x) + (1-\kappa)x_u$ and take the infimum on the resulting inequality over $x \in X$. □

From item (iii) of Corollary 3, we know that every accumulation point of $\overline{x}^\varepsilon$ is a global solution of (39). In practice, the result applies because it is possible to compute $\overline{x}^\varepsilon$ as global solutions and, for this setting, smoothing preserves the original convexity of the problem. In general, for non-convex value functions, item (iii) is still true, but one may not be sure of global optimality in computation. Whenever gradient consistency holds, limits of $\overline{x}^\varepsilon$ are stationary points of the original problem. In particular, the local boundedness of the smoothed gradients proved in this paper ensures that the singular subdifferential of the value function and the singular closure of the smoothed gradients agree, which is the gradient consistency result.

7 Numerical experiments

We now benchmark our proposal against the state-of-the-art bundle solver [Fra02] in terms of decrease in the objective function values along the iterations, using data profiles [MW09].

The experiments were performed on an Intel Core i7 computer with 1.9 GHz, 8 cores and 15.5 GB RAM, running under Ubuntu 18.04.3 LTS.

7.1 Instances and solvers considered in the benchmark

The test set was created by using four functions from I. Deák's collection [Deá06], having $nx = 20$ first-stage variables and $ny = 30$ second-stage variables per scenario. For each scenario, $l = 20$ affine equality constraints

couple the two stages and there are 10 affine equality constraints defining the fist-stage feasible set X. All the assumptions in Corollary 3 are satisfied.

In order to define new, more challenging, instances, the stochastic linear programs from [Deá06] were modified by adding a quadratic term to the linear second-stage cost. Accordingly, given $\mathfrak{q}_s$, the linear cost in the original problems and a scalar parameter $r \geq 0$, in (39) we set

$$q_s(y) := \mathfrak{q}_s^\top y + \frac{1}{2} r y^\top y.$$

The instances in the benchmark are obtained by varying the number of scenarios and the quadratic parameter

$$S \in \{1,2,\ldots,20\} \quad \text{and} \quad r \in \{0, 0.01, 0.1, 1\}.$$

In (39) the risk-aversion parameter is $\kappa \in \{0.5, 1\}$, and the confidence level is set to $\alpha = 0.9$, noting that the risk-neutral version ($\kappa = 1$) has no variables x_u, z and related constraints. Accordingly, the considered second-stage problems are increasingly more difficult, being linear programs if $r = 0$ and $\kappa = 1$, quadratic programs if $r > 0$ and $\kappa = 1$, and problems with quadratic objective and quadratic constraints (QCQP) if $r > 0$ and $\kappa \in [0, 1)$. We used CPLEX 12.8 and an optimized build of Ipopt 3.12.10 with the linear solver Pardiso as described in the manual; see also [WLB06]. Both packages were configured to employ only one thread per run.

To solve the corresponding problems (39), we consider two methodologies, listed below.

– BM, a decomposition method for the first-stage problem, based on the bundle algorithm by A. Frangioni, [Fra02], one of the best solvers in the area. The method parameters were tuned for best performance, particularly regarding the management of the bundle size (keeping only active bundle elements). At each iteration, say (x^k, x_u^k), the algorithm uses certain oracle information, obtained by evaluating the nondifferentiable convex objective function

$$c(x^k) + (1-\kappa)x_u^k + \sum_{s=1}^{S} p_s \mathcal{Q}_s(x^k, x_u^k).$$

In addition to this value, the bundle method uses a subgradient of the form

$$\left(\nabla c(x^k) + \sum_{s=1}^{S} p_s T_s^\top \eta_{x^k}, (1-\kappa) - \sum_{s=1}^{S} p_s \eta_{x_u^k}\right)^\top,$$

for multipliers $(\eta_{x^k}, \eta_{x_u^k})$ obtained when computing the value of the recourse function $\mathcal{Q}_s(x^k, x_u^k)$, for each scenario s. Depending on the instance, computing such value amounts to dealing with a linear program, a quadratic program, or a QCQP problem, solved with the packages CPLEX or Ipopt. As CPLEX currently does not provide directly multipliers for quadratic constraints, we could not use it for the risk-averse quadratic runs.

– ST, our smoothing with log-barrier and Tikhonov regularization approach, solving the approximate first-stage master problem (40) with Ipopt. In this setting, the oracle information for the smoothed objective function

$$c(x^k) + (1-\kappa)x_u^k + \sum_{s=1}^{S} p_s \left(\kappa q_s(y^\varepsilon(x^k, x_u^k)) + \frac{1-\kappa}{1-\alpha} z^\varepsilon(x^k, x_u^k)\right)$$

requires the solution of one problem (40) written with $(x, x_u) = (x^k, x_u^k)$ per scenario s. For all the considered instances, this is a problem with nonlinear objective function and affine constraints solved with Ipopt, giving the objective function gradient as callback information, computing its value according to Theorem 1(ii). The performance reported below relies heavily on the availability of an optimized build of Ipopt. In particular, the regularized solution mappings in Theorem 1(i), $y^\varepsilon(x)$ and $\lambda^\varepsilon(x)$, are an output of Ipopt, once certain *mu-target* option is activated (such Ipopt parameter corresponds to ε). For simplicity, ε was kept constant along iterations. However, note that the bounds given in item (i) in Corollary 3 justify interpreting this parameter as a direct measure of precision when $\mu = 0$. Recall that when $\mu > 0$, as shown in Proposition 2 and further discussed in Remark 1, the determination of the quality of the smoothing depends on bounds for the scenario subproblem solutions, a knowledge that is hardly available in practice. For numerically hard problems taking $\mu > 0$ can be advantageous to improve the chances that derivatives of the regularized solution mappings are sufficiently precise. As in (8), in this case the Tikhonov term

involves a factor $0.5\varepsilon\mu$ that is kept constant along iterations. The range chosen for these parameters is

$$\varepsilon \in \{0.01, 0.1, 1\} \quad \text{and} \quad \mu \in \{0, 0.1, 1\}.$$

As we deal with random instances, each experiment is repeated three times, yielding 540 or 2160 different runs, respectively if $r = 0$ and $r > 0$. Table 1 summarizes all the variants considered in the benchmark.

Problem type	(in (39))	BM-CPLEX	BM-Ipopt	ST-Ipopt
Risk-neutral linear	$(\kappa = 1, r = 0)$	x	x	x
Risk-neutral quadratic	$(\kappa = 1, r > 0)$	x	x	x
Risk-averse linear	$(\kappa \in [0,1), r = 0)$	–	x	x
Risk-averse quadratic	$(\kappa \in [0,1), r > 0)$	–	x	x

Table 1 Benchmark configuration.

7.2 Comparing the solvers with data profiles

To report the results of the experiments we use data profiles as introduced in [MW09]. Specifically, for a given instance, the maximum running time of a given set of methods is used to normalize all the running times, so that in the graph abscissa the range for all methods is between 0 and 1 (the value of 1 can be thought of as the maximum time budget given to the solvers). The ordinate in the data profiles corresponds to the probability of each method delivering the best iterate plus a gap until a time given in the abscissa. The gap corresponds to 5% of the largest decrease obtained for a given instance by all methods, that were given the same starting point.

The results are analyzed by considering the different groups in Table 1, starting with the risk-neutral instances ($\kappa = 1$), in both its linear ($r = 0$) and quadratic ($r > 0$) variants. The corresponding profiles are given in Figure 1.

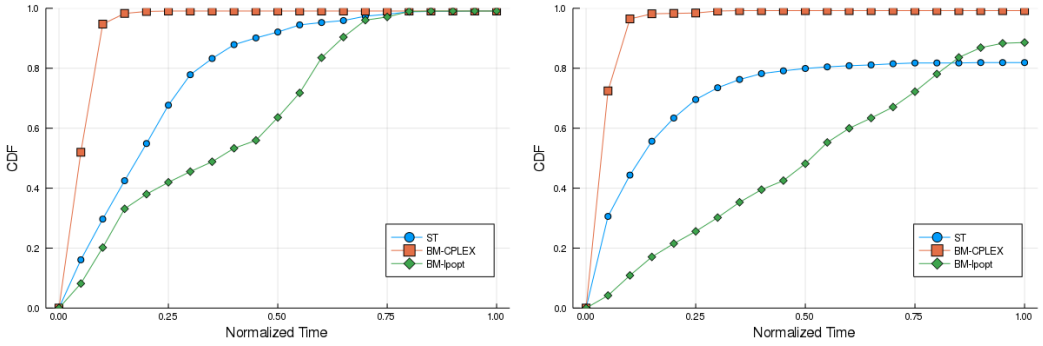

Fig. 1 Performance for linear (left) and quadratic (right) instances without risk.

In both graphs BM-CPLEX is a clear winner, followed by ST and with BM-Ipopt performing worst. For the linear group, in 70% of the runs ST obtained the largest functional decrease using a slightly more than a quarter of the time budget: on the left graph the abscissa 0.25 has ordinate 0.7 for ST (the dot). All the solvers succeeded in solving all of the linear instances (the ordinate value of 1 is attained by the three lines). By contrast, the quadratic instances clearly put Ipopt in trouble, as both ST and BM-Ipopt failed in about 20% of the runs. Notice that for this simplest test set (no risk) there is a big difference in the performance of BM-CPLEX and BM-Ipopt. This illustrates well the impact that subproblem solution times can have on a decomposition method. Considering that BM subproblems are all linear or quadratic programs for these groups of instances, the profiles can be seen as a cautionary tale on the importance of using a specific solver (CPLEX) rather than a general purpose one (Ipopt) whenever possible. Incidentally, this behavior also indicates that the difference of performance between ST and BM-CPLEX might be explained by the time each solver spent in the respective subproblems (nonlinear for ST).

The next profiles in Figure 2, potentially more challenging in terms of subproblem solution, consider risk aversion ($\kappa \in [0,1)$), again with linear ($r = 0$) instances on the left and quadratic ones ($r > 0$) on the right.

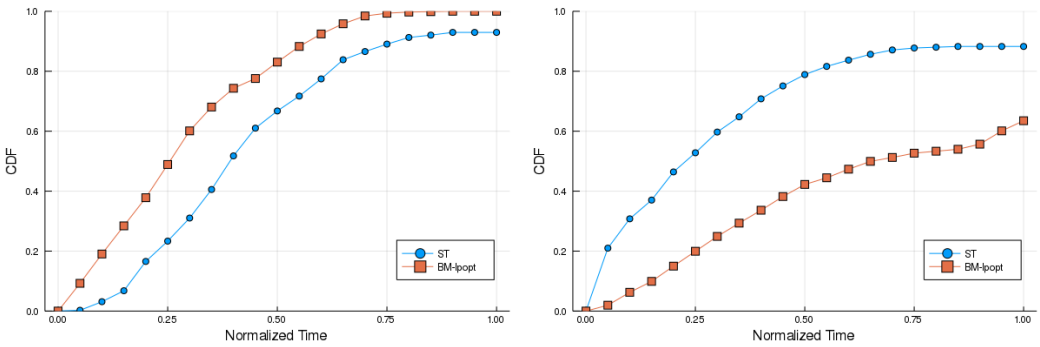

Fig. 2 Performance for linear (left) and quadratic (right) instances with risk.

The left graph gives BM-Ipopt as a winner, followed closely by ST. The situation is reversed for the risk-averse quadratic set of instances. Specifically, on the right graph ST performance is far superior than BM-Ipopt's (recall that for these instances the comparison with BM-CPLEX is not possible). The fact that these are the hardest problems is evident in the profile on the right, showing a percentage of failures of 10% and 40%, for ST and BM-Ipopt, respectively.

In our final profiles we confirm the impact in terms of solution times of introducing a quadratic term in the second-stage subproblems, particularly when there is risk aversion. The top profiles in Figure 3 show that, when r varies in $\{0.01, 0.1, 1.0\}$ both BM-CPLEX and ST (left and right top graphs) perform alike for the instances without risk aversion. The situation is substantially different for the bottom profiles, with the performance of BM-Ipopt and ST for the same three values of r, now considering risk. On the left bottom graph, as r gets smaller, the improvement in BM-Ipopt's performance is noticeable, as well as a reduction in the percentage of failures: about 30% for $r = 0.01$ and $r = 0.1$ and 60% for $r = 1.0$. For both BM and ST, smaller values of r make the problem solution easier. The right bottom graph, with ST runs, has much less failures than BM's, and, more remarkably, the three ST lines look alike for the three different values of r.

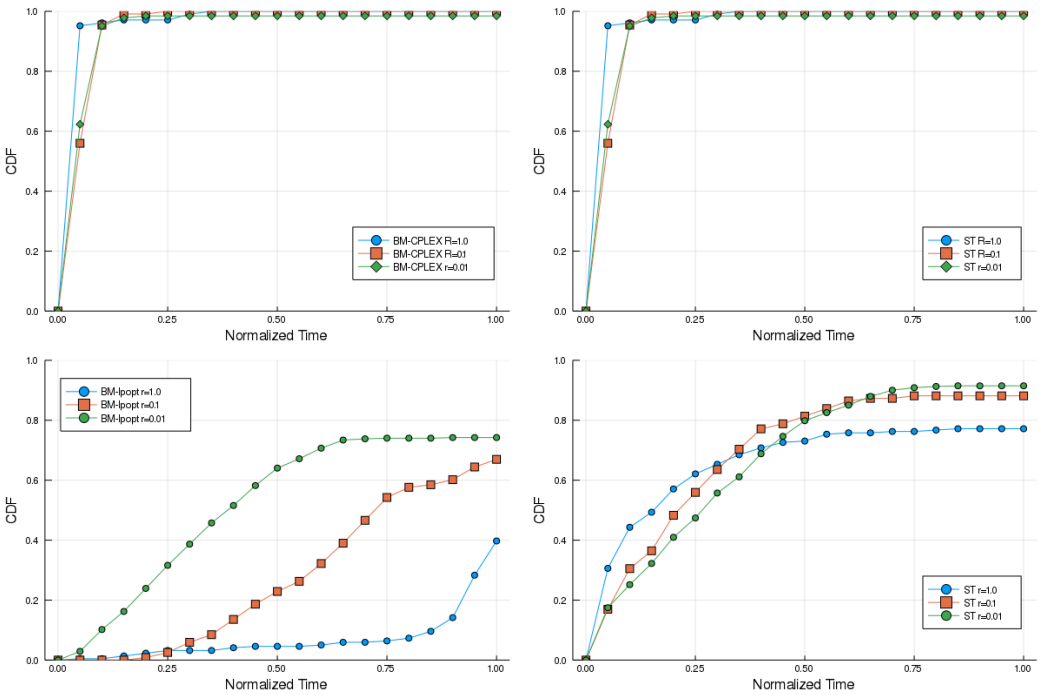

Fig. 3 Effect of the quadratic term on BM (left) and ST (right) without (top) and with (bottom) risk.

For the considered test set, it appears that BM-CPLEX should be preferred for the linear instances, while ST is the winner for problems with risk aversion and quadratic objective function in the second stage. Regarding ST's failures, by solving the deterministic equivalent with CPLEX, we could check that ST had found good estimates of the optimal value without reaching the threshold of 5% error in some instances. We expect that a dynamical management of ε would help in eliminating such failures. This is a topic of future research.

Acknowledgments The authors thank the referees and Editor for beneficial comments. The first and second authors are grateful to Ecole Polytechnique, France, for the support through the 2018-2019 Gaspard Monge Visiting Professor Program. Research of the second author is partly funded by CNPq Grant 306089/2019-0, CEPID CeMEAI, and FAPERJ in Brazil. The third author is supported by CNPq Grant 303913/2019-3, by FAPERJ Grant E-26/202.540/2019, by PRONEX–Optimization, and by the Russian Foundation for Basic Research grant 19-51-12003 NNIOa.

References

[Ahm06] S. Ahmed. "Convexity and decomposition of mean-risk stochastic programs". *Mathematical Programming* 106.3 (2006), pp. 433–446.

[Att77] H. Attouch. "Convergence de fonctions convexes, de sous-differentiels et semi-groupes". *Comptes Rendus de l'Academie des Sciences de Paris* 284.1 (1977), pp. 539–542.

[BGKKT83] B. Bank, J. Guddat, D. Klatte, B. Kummer, and K. Tammer. *Non-Linear Parametric Optimization*. Springer, 1983.

[BGLS06] J. Bonnans, J. Gilbert, C. Lemaréchal, and C. Sagastizábal. *Numerical Optimization. Theoretical and Practical Aspects*. Universitext. Berlin: Springer-Verlag, 2006.

[BH13] J. Burke and T. Hoheisel. "Epi-convergent Smoothing with Applications to Convex Composite Functions". *SIAM Journal on Optimization* 23.3 (2013), pp. 1457–1479.

[BH17] J. Burke and T. Hoheisel. "Epi-Convergence Properties of Smoothing by Infimal Convolution". *Set-Valued and Variational Analysis* 25.1 (2017), pp. 1–23.

[BHK13] J. Burke, T. Hoheisel, and C. Kanzow. "Gradient Consistency for Integral-Convolution Smoothing Functions". *Set-Valued and Variational Analysis* 21.2 (2013), pp. 359–376.

[BS00] J. F. Bonnans and A. Shapiro. *Perturbation Analysis Of Optimization Problems*. Springer, 2000.

[BT12] A. Beck and M. Teboulle. "Smoothing and First Order Methods: A Unified Framework". *SIAM Journal on Optimization* 22.2 (2012), pp. 557–580.

[Che12] X. Chen. "Smoothing Methods for Nonsmooth, Nonconvex Minimization". *Mathematical Programming* 134.1 (2012), pp. 71–99.

[CQS98] X. Chen, L. Qi, and D. Sun. "Global and Superlinear Convergence of the Smoothing Newton Method and Its Application to General Box Constrained Variational Inequalities". *Mathematics of Computation* 67.222 (1998), pp. 519–540.

[Deá06] I. Deák. "Two-stage Stochastic Problems with Correlated Normal Variables: Computational Experiences". *Annals of Operations Research* 142.1 (2006), pp. 79–97.

[DGL12] N. Dinh, M. Goberna, and M. López. "On the stability of the optimal value and the optimal set in optimization problems". *Journal of Convex Analysis* 19 (2012), pp. 927–953.

[DM15] S. Dempe and P. Mehlitz. "Lipschitz Continuity of the Optimal Value Function in Parametric Optimization". *Journal of Global Optimization* 61.2 (2015), pp. 363–377.

[DRS09] D. Dentcheva, A. Ruszczyński, and A. Shapiro. *Lectures on Stochastic Programming*. SIAM, Philadelphia, 2009.

[DS99] L. Drummond and B. Svaiter. "On well definedness of the Central Path". *Journal of Optimization Theory and Applications* 102.2 (1999), pp. 223–237.

[FI90] A. V. Fiacco and Y. Ishizuka. "Sensitivity and stability analysis for nonlinear programming". *Annals of Operations Research* 27.1 (1990), pp. 215–235.

[Fia83] A. V. Fiacco. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming*. 1983.

[FM68] A. V. Fiacco and G. P. McCormick. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. Wiley, 1968.

[Fra02] A. Frangioni. "Generalized Bundle Methods". *SIAM Journal on Optimization* 13.1 (2002), pp. 117–156.

[GLYZ14] L. Guo, G.-H. Lin, J. Ye, and J. Zhang. "Sensitivity Analysis of the Value Function for Parametric Mathematical Programs with Equilibrium Constraints". *SIAM Journal on Optimization* 24.3 (2014), pp. 1206–1237.

[HBT18] L. Hellemo, P. Barton, and A. Tomasgard. "Decision-dependent probabilities in stochastic programs with recourse". *Computational Management Science* 15.3 (2018), pp. 369–395.

[IS06] A. Izmailov and M. Solodov. "A Note on Error Estimates for some Interior Penalty Methods". *Recent Advances in Optimization. Lecture Notes in Economics and Mathematical Systems* 563 (2006).

[Lan93] S. Lang. *Real and Functional Analysis*. Graduate Texts in Mathematics, Springer, 1993.

[LCPS18] J. Liu, Y. Cui, J.-S. Pang, and S. Sen. *Two-stage Stochastic Programming with Linearly Bi-parameterized Quadratic Recourse*. Tech. rep. The University of Southern California, December/2018, 2018.

[MNY09] B. S. Mordukhovich, N. M. Nam, and N. D. Yen. "Subgradients of Marginal Functions in Parametric Mathematical Programming". *Mathematical Programming* 116.1-2 (2009), pp. 369–396.

[Mor06] B. S. Mordukhovich. *Variational Analysis and Generalized Differentiation I*. Springer Berlin Heidelberg, 2006.

[Mor18] B. S. Mordukhovich. *Variational Analysis and Applications*. Springer International Publishing, 2018.

[MW09] J. J. Moré and S. M. Wild. "Benchmarking Derivative-Free Optimization Algorithms". *SIAM Journal on Optimization* 20.1 (2009), pp. 172–191.

[MZ98] R. D. Monteiro and F. Zhou. "On the Existence and Convergence of the Central Path for Convex Programming and Some Duality Results". *Computational Optimization and Applications* 10.1 (1998), pp. 51–77.

[Nes05] Y. Nesterov. "Smooth Minimization of Non-Smooth Functions". *Mathematical Programming* 103.1 (2005), pp. 127–152.

[OS14] W. Oliveira and C. Sagastizábal. "Level bundle methods for oracles with on-demand accuracy". *Optimization Methods and Software* 29.6 (2014), pp. 1180–1209.

[OSL14] W. Oliveira, C. Sagastizábal, and C. Lemaréchal. "Convex proximal bundle methods in depth: a unified analysis for inexact oracles". *Mathematical Programming* 148.1-2 (2014), pp. 241–277.

[QSZ00] L. Qi, D. Sun, and G. Zhou. "A new look at smoothing Newton methods for nonlinear complementarity problems and box constrained variational inequalities". *Mathematical Programming* 87.1 (2000), pp. 1–35.

[RU02] R. Rockafellar and S. Uryasev. "Conditional Value-at-Risk for General Loss Distributions". *Journal of Banking and Finance* 26.7 (2002), pp. 1443–1471.

[RW09] T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer, 2009.

[RX05] D. Ralph and H. Xu. "Implicit Smoothing and Its Application to Optimization with Piecewise Smooth Equality Constraints". *Journal of Optimization Theory and Applications* 124.3 (2005), pp. 673–699.

[Sag12] C. Sagastizábal. "Divide to conquer: decomposition methods for energy optimization". *Mathematical Programming* 134.1 (2012), pp. 187–222.

[VW69] R. Van Slyke and R. Wets. "L-shaped linear programs with applications to control and stochastic programming". *SIAM Journal on Applied Mathematics* 17 (1969), pp. 638–663.

[WLB06] A. Wachter and T. L. Biegler. "On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming". *Mathematical Programming* 106.1 (2006), pp. 25–57.

[Wri97] S. J. Wright. *Primal-Dual Interior-Point Methods*. Society for Industrial and Applied Mathematics Philadelphia, PA, USA, 1997.