SIAM J. OPTIM.
Vol. 30, No. 3, pp. 1996–2025

AN APPROXIMATION SCHEME FOR DISTRIBUTIONALLY ROBUST NONLINEAR OPTIMIZATION*

JOHANNES MILZ† AND MICHAEL ULBRICH†

Abstract. We consider distributionally robust optimization problems (DROPs) with nonlinear and nonconcave dependence on uncertain parameters. The DROP can be written as a nonsmooth, nonlinear program with a bilevel structure; the objective function and each of the constraint functions are suprema of expected values of parametric functions taken over an ambiguity set of probability distributions. We define ambiguity sets through moment constraints, and to make the computation of first order stationary points tractable, we approximate nonlinear functions using quadratic expansions w.r.t. parameters, resulting in lower-level problems defined by trust-region problems and semidefinite programs. Subsequently, we construct smoothing functions for the approximate lower level functions which are computationally tractable, employing strong duality for trust-region problems, and show that gradient consistency holds. We formulate smoothed DROPs and apply a homotopy method that dynamically decreases smoothing parameters and establish its convergence to stationary points of the approximate DROP under mild assumptions. Through our scheme, we provide a new approach to robust nonlinear optimization as well. We perform numerical experiments and comparisons to other methods on a well-known test set, assuming design variables are subject to implementation errors, which provides a representative set of numerical examples.

Key words. distributionally robust optimization, robust optimization, trust-region problem, semidefinite programming, smoothing functions, gradient consistency, smoothing methods

AMS subject classifications. 90C26, 90C30, 90C46, 90C59, 49M37

DOI. 10.1137/19M1263121

1. Introduction.

We develop an approximation scheme for the nonlinear distributionally robust optimization problem (DROP)

$$\min_{x\in X} \sup_{P\in\mathcal{P}} \mathbb{E}_P[f_0(x,\xi)] \quad \text{s.t.} \quad \sup_{P\in\mathcal{P}} \mathbb{E}_P[f_j(x,\xi)] \le 0, \quad j \in J\setminus\{0\}, \tag{1.1}$$

where $X \subset \mathbb{R}^n$ is the set of design variables and $f_j : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$, $j \in J \subset \mathbb{N}_0$. The ambiguity set $\mathcal{P}$ is defined through moment constraints of the random vector ξ and entropic dominance similar to [20, 23, 55]:

$$\begin{aligned}\mathcal{P} = \{\, P \in \mathcal{M} : \|\bar\Sigma^{-\frac{1}{2}}(\mathbb{E}_P[\xi]-\bar\mu)\|_2 \le \Delta, \quad \bar\Sigma_0 \preccurlyeq \mathrm{Cov}_P[\xi] \preccurlyeq \bar\Sigma_1,\\ \ln \mathbb{E}_P[\exp(y^T(\xi-\mathbb{E}_P[\xi]))] \le (1/2)y^T\bar\Sigma_1 y \quad \text{for all} \quad y\in\mathbb{R}^p \,\},\end{aligned} \tag{1.2}$$

where $\Delta > 0$, $\bar\mu \in \mathbb{R}^p$, and $\bar\Sigma_0$, $\bar\Sigma_1$, $\bar\Sigma \in \mathbb{R}^{p\times p}$ are symmetric, $\bar\Sigma_0$, $\bar\Sigma_1$, and $\bar\Sigma_1 - \bar\Sigma_0$ are positive semidefinite, and $\bar\Sigma$ is positive definite. Moreover, $\mathcal{M}$ denotes the set of probability distributions of ξ on $\mathbb{R}^p$.

To obtain tractable approximations of the objective and constraint functions of (1.1), we approximate $f_j(x,\cdot)$ using second order expansions $m_j(x,\cdot)$ defined by

$$m_j(x,\xi) = a_j(x) + b_j(x)^T(\xi-\bar\mu) + (1/2)(\xi-\bar\mu)^T C_j(x)(\xi-\bar\mu), \tag{1.3}$$

*Received by the editors May 21, 2019; accepted for publication (in revised form) May 19, 2020; published electronically July 30, 2020.
https://doi.org/10.1137/19M1263121
Funding: The authors were supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project 188264188/GRK1754, through the International Research Training Group IGDK 1754, "Optimization and Numerical Analysis for Partial Differential Equations with Nonsmooth Structures."
†Chair of Mathematical Optimization, Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany (milz@ma.tum.de, mulbrich@ma.tum.de).

where $a_j : \mathbb{R}^n \to \mathbb{R}$, $b_j : \mathbb{R}^n \to \mathbb{R}^p$, $C_j : \mathbb{R}^n \to \mathbb{S}^p$. We formulate the approximated DROP

$$\min_{x\in X} \sup_{P\in\mathcal{P}} \mathbb{E}_P[m_0(x,\xi)] \quad \text{s.t.} \quad \sup_{P\in\mathcal{P}} \mathbb{E}_P[m_j(x,\xi)] \le 0, \quad j \in J \setminus \{0\}. \tag{1.4}$$

The definition of the ambiguity set $\mathcal{P}$ (see (1.2)) and

$$\mathbb{E}_P[m_j(x,\xi)] = a_j(x) + b_j(x)^T d + (1/2)d^T C_j(x) d + (1/2) C_j(x) \bullet \Sigma,$$

where $d = \mathbb{E}_P[\xi] - \bar{\mu}$ and $\Sigma = \mathrm{Cov}_P[\xi]$, imply that each lower-level optimization problem of (1.4) separates into the semidefinite program (SDP)

$$\varphi_j(x) = \max_{\Sigma\in\mathbb{S}^p} \left\{ (1/2) C_j(x) \bullet \Sigma : \quad \bar{\Sigma}_0 \preccurlyeq \Sigma \preccurlyeq \bar{\Sigma}_1 \right\} \tag{1.5}$$

and the nonconvex trust-region problem (TRP)

$$\psi_j(x) = a_j(x) + \max_{d\in\mathbb{R}^p} \left\{ b_j(x)^T d + (1/2) d^T C_j(x) d : \quad \|\bar{\Sigma}^{-1/2} d\|_2 \le \Delta \right\}, \tag{1.6}$$

where $\psi_j : \mathbb{R}^n \to \mathbb{R}$ and $\varphi_j : \mathbb{R}^n \to \mathbb{R}$.

The optimal value functions (1.5) and (1.6) provide a tractable approximation of the lower-level problems in (1.1). These functions lack higher order differentiability, motivating us to construct smoothing functions for them. We propose a homotopy method similar to the smoothing methods in [18, 64] to solve a sequence of smoothed DROPs to obtain a Clarke stationary point of the approximated DROP (1.4).

The SDP in (1.5) can be solved analytically after computing the eigenvalues of a transformation of $C_j(x)$; see [65, Thm. 2.2]. We make use of this and apply results on spectral functions, such as statements established in [41, 60], to obtain a smoothing function of (1.5). Our approach for the value function of the TRP (1.6) utilizes strong duality for TRPs; see, e.g., [58]. We apply a reciprocal barrier function to its dual and observe that the dual is equivalent to a TRP.

Distributionally robust optimization (DRO) is a popular methodology used to obtain robust solutions to optimization problems under uncertainty; cf. [23, 26, 31, 53, 63]. It "robustifies" against distributions contained in an ambiguity set. If this set is a singleton, DRO is reduced to stochastic optimization; see [54]. A very popular choice for constructing an ambiguity set is based on moment constraints of the parameters, such as the one in (1.2); cf. [23, 54, 55, 63]. Another approach is to define the set by measures close to a reference measure w.r.t. a certain distance; cf. [30, 53, 68].

Some special classes of DROPs can be transformed into one-level problems using Lagrangian duality. For example, if ambiguity sets are conic representable, maximization problems w.r.t. probability measures become conic linear programs and, therefore, can be transformed into minimization problems and concatenated with upper-level problems; cf. [23]. If suitable assumptions, such as the convexity of the objective function w.r.t. design variables, are satisfied, the resulting optimization problem is tractable [23, 63]. The reformulation of the lower-level problems of (1.4) as linear matrix inequalities has been discussed in the supplementary material of [63].

If the SDP (1.5) is removed from (1.4), we obtain the robust optimization problem (ROP)

$$\min_{x\in X} \psi_0(x) \quad \text{s.t.} \quad \psi_j(x) \le 0, \quad j \in J \setminus \{0\}. \tag{1.7}$$

Research on robust optimization (RO) may be divided into contributions assuming concave dependence w.r.t. parameters (see, e.g., [2, 3, 5, 7]) and those assuming nonconcave dependence (see e.g., [24, 34, 67]). The authors of [24] and [67] use a linearization scheme for nonlinear RO to obtain tractable approximations of lower-level problems, resulting in a nonlinear second order cone program if an ellipsoidal uncertainty set is used. Instead of linearization, second order models are applied in [38, 39]. These expansions may be more effective than linearizations and may provide a trade-off between accuracy and tractability; cf. [34, 39]. This approach results in constraints such as the one in (1.7), which are reformulated using its canonical, necessary and sufficient optimality conditions in [38, 39]. The resulting problem is a mathematical program with complementarity constraints (MPCC); see, e.g., [35, 57]. In addition, the constraint set contains linear matrix inequalities, requiring the Hessian matrix of the Lagrangian of each robustified constraint to be positive semidefinite. In [38, 39] the inequalities are reformulated using eigenvalue constraints, introducing nonsmooth constraint functions. Moreover, in [34] a numerical scheme for nonlinear min-max optimization problems was developed. Nonconvex ROPs without approximation schemes were considered in, e.g., [8, 9]. The lower-level problems in (1.7) may be reformulated as SDPs; see [3, sect. 1.4 and Lem. 14.37].

Using Lagrangian duality for both (1.5) and (1.6) (see, e.g., [5, Chap. 4] and [12, sect. B.1]), we can show that, for $\rho \in \mathbb{R}$, $x \in X$, $\Delta > 0$, and $\bar{\Sigma}_1 - \bar{\Sigma}_0$ positive definite, the condition $\psi_j(x) + \varphi_j(x) \le \rho$ is satisfied if and only if there exists $(\gamma_j, \lambda_j, \Lambda_j, \Upsilon_j) \in \mathbb{R} \times \mathbb{R}_+ \times \mathbb{S}^p_+ \times \mathbb{S}^p_+$ such that

$$
\begin{aligned}
&2a_j(x) - \gamma_j - (\bar{\Sigma}^{-1/2}\bar{\Sigma}_0\bar{\Sigma}^{-1/2}) \bullet \Lambda_j + (\bar{\Sigma}^{-1/2}\bar{\Sigma}_1\bar{\Sigma}^{-1/2}) \bullet \Upsilon_j \le 2\rho,\\
&\begin{bmatrix} \lambda_j I - \bar{\Sigma}^{1/2} C_j(x) \bar{\Sigma}^{1/2} & -\bar{\Sigma}^{1/2} b_j(x) \\ -(\bar{\Sigma}^{1/2} b_j(x))^T & -\lambda_j \Delta^2 - \gamma_j \end{bmatrix} \succcurlyeq 0, \quad \Upsilon_j - \Lambda_j = -\bar{\Sigma}^{1/2} C_j(x) \bar{\Sigma}^{1/2}.
\end{aligned} \tag{1.8}
$$

Hence, (1.4) can be reformulated as a nonlinear SDP (NSDP). We refer the reader to [66] for a survey for optimization methods for NSDPs. Computer codes, such as `PENLAB` [28], require first derivatives of the constraints (1.8). Our approach allows the numerical treatment of (1.4) via a sequence of standard nonlinear programs (NLPs). Derivatives required for each NLP may be easier to obtain than those of an NSDP formulation. In particular, our approach requires the derivative of $\mathbb{R}^n \ni x \mapsto d^T C_j(x) d$, $d \in \mathbb{R}^p$, rather than of the mapping $C_j : \mathbb{R}^n \to \mathbb{S}^p$.

As a further alternative, algorithms for nonsmooth, nonconvex optimization can be applied to (1.4). Different algorithms, such as subgradient and bundle methods, for this problem class are compared in [37]. Further methods include gradient sampling algorithms [17] and quasi-Newton methods [42]. Bundle methods, such as `MPBNGC` [43, 44], when applied to (1.4), require evaluations of the objective and constraint functions of (1.4) as well as subgradients.

The computational cost of evaluating the smoothing function of $\varphi_j + \psi_j$ and its gradient is similar to that of $\varphi_j + \psi_j$ and of one of its subgradients; cf. section 7.

We compare our algorithmic scheme with the proximal bundle method `MPBNGC` applied to (1.4) and `PENLAB` applied to an NSDP reformulation of (1.4) in section 7. `MPBNGC` and `PENLAB` are both open-source. The decision tree for nonsmooth optimization software, `Solver-o-matic` [36], recommended the use of `MPBNGC` as a solver for the nonsmooth, nonconvex optimization problem (1.4). `Solver-o-matic` includes eight solvers for nonsmooth, nonconvex optimization. The comparison of nonsmooth minimization methods made in [37] indicates that `MPBNGC` is an efficient solver for nonsmooth optimization problems.

Smoothing methods are popular schemes for the solution of nonconvex, nonsmooth, and Lipschitz optimization problems; see, e.g., [14, 18, 64]. Our algorithmic scheme is related to recent contributions, such as [14, 15, 18, 64], in that it provides further examples of smoothing functions and applies their concepts and methodology. We apply an NLP solver to compute stationary points of a sequence of smoothed DROPs generated by the decreasing parameters and, therefore, our algorithmic approach is similar to those of [18, 64]. For our numerical experiments, we use the open-source, state-of-the-art solver `Ipopt` [61] to compute approximate stationary points of these NLPs. However, any solver for nonconvex, nonlinear programming can be used.

Our scheme relies on an approximation of the lower-level problems in (1.1). However, we are able to compute stationary points of the approximation (1.4) of (1.1) without the assumption that computationally available bounds on the Hessian matrix of $f_j(x,\cdot)$ as in [34] are known, and we do not require expensive numerical schemes as in [8, 9]. Our reformulation does not result in an MPCC or an NSDP, and we do not increase the dimension of the initial DROP or ROP. A further advantage is that we obtain standard NLPs with tractable objective and constraints. These conditions are all favorable from a computational point of view because, e.g., an implementation of further algorithms is not required, making our approach applicable to many problems.

Outline of the paper. In section 2, the choice of the ambiguity set $\mathcal{P}$ (see (1.2)) is explained. Section 3 introduces the concept of smoothing functions, a smoothed DROP of (1.4), and a homotopy method used for the numerical solution of (1.4). Section 4 presents our smoothing approach for the SDPs in (1.5), which utilizes the theory of spectral functions. In section 5, our smoothing scheme for the TRPs in (1.6) is presented. It is based on the strong duality of TRPs. Global convergence of the homotopy method is shown in section 6. Section 7 presents numerical examples illustrating that the approximation scheme (1.4) of (1.1) can be effective. Section 8 presents a concise summary of our contributions.

Notation. The set of symmetric $m \times m$ matrices is $\mathbb{S}^m$. We refer to $\mathbb{S}^m_{++} \subset \mathbb{S}^m$ ($\mathbb{S}^m_+ \subset \mathbb{S}^m$) as the set of positive (semi)definite matrices. The identity matrix is I. The eigenvalue mapping is $\lambda : \mathbb{S}^p \to \mathbb{R}^p$, where $\lambda(A)$ contains the eigenvalues of A in decreasing order, i.e., $\lambda_{\max}(A) = \lambda_1(A) \geq \cdots \geq \lambda_p(A) = \lambda_{\min}(A)$. Here, $A \succcurlyeq B$ ($A \succ B$) for $A, B \in \mathbb{S}^m$ means $A - B \in \mathbb{S}^m_+$ ($A - B \in \mathbb{S}^m_{++}$). We use $\bullet$ to denote the Frobenius inner-product on $\mathbb{S}^m$. The set $N(A)$ is the null space of $A \in \mathbb{S}^m$. The matrix $A^{1/2} \in \mathbb{S}^m$ is the unique symmetric square root of $A \in \mathbb{S}^m_+$, B^+ is the Moore–Penrose inverse of $B \in \mathbb{R}^{m\times m}$, $|J|$ is the cardinality of the set J, and $(\cdot)_+ = \max\{0, \cdot\}$. For $a \in \mathbb{R}^m$, $\mathrm{Diag}(a) \in \mathbb{S}^m$ is the diagonal matrix with $(\mathrm{Diag}(a))_{ii} = a_i$. The Euclidean norm ($\infty$-norm) on $\mathbb{R}^m$ is $\|\cdot\|_2$ ($\|\cdot\|_\infty$). The convex hull of $A \subset \mathbb{R}^{n\times m}$ is $\mathrm{conv}\, A$. A function $h : \mathbb{R}^m \to \mathbb{R}$ is symmetric if it is invariant under coordinate permutations; see, e.g., [40]. The gradient of $G : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ w.r.t. x evaluated at (x, y) is denoted by $\nabla_x G(x, y)$. For $A : \mathbb{R}^n \to \mathbb{S}^p$, we denote by $DA(x)$ its derivative and by $DA(x)^*$ its adjoint operator evaluated at $x \in \mathbb{R}^n$. The set $\partial G(y)$ is the Clarke subdifferential of $G : \mathbb{R}^n \to \mathbb{R}$ at $y \in \mathbb{R}^n$ (cf. [21, p. 27]) consisting of column vectors. We use $\mathbb{E}_P[\xi]$ and $\mathrm{Cov}_P[\xi]$ to denote the mean and covariance of ξ w.r.t. $P \in \mathcal{M}$, respectively. Here, $\mathcal{M}$ is the set of probability distributions of ξ on $\mathbb{R}^p$. The normal distribution with mean $\mu \in \mathbb{R}^p$ and covariance matrix $\Sigma \in \mathbb{S}^p_+$ is $N(\mu, \Sigma)$.

2. Choice of ambiguity set. We comment on the choice of the ambiguity set $\mathcal{P}$ defined in (1.2), discuss conditions implying that the objective and constraint

Algorithm 3.1. Homotopy method.

Choose parameters $t_0 \in \mathbb{R}^3_{++}$, $t_{\min} \in \mathbb{R}^3_{+}$, $\varepsilon_0 > 0$, $\varepsilon_{\min} \geq 0$, and $\rho \in (0, 1)$.
For $k = 0, 1, \ldots$

1. Compute an ε_k-KKT-tuple (x^k, ϑ^k) of (3.2) for $t = t^k$.
2. If $t^k \leq t_{\min}$ and $\varepsilon_k \leq \varepsilon_{\min}$ hold, STOP and return (x^k, ϑ^k).
3. Compute $0 < t^{k+1} \leq \rho t^k$ and $\varepsilon_{k+1} = \rho \varepsilon_k$.

functions of the DROPs (1.1) and (1.4) are finite-valued, and suggest choices to define the quadratic model functions m_j (see (1.3)) of f_j.

The first two conditions on $\mathbb{E}_P[\xi]$ and $\mathrm{Cov}_P[\xi]$ imposed by $\mathcal{P}$ (see (1.2)) model confidence regions of the mean and the covariance of ξ under suitable assumptions, respectively. In a data-driven framework, the data defining the ambiguity set (1.2) may be chosen similarly to the choices in [23, 55]. It can be shown that $\sup_{P \in \mathcal{P}} \mathbb{E}_P[\|\xi\|_2^\gamma] < \infty$ for all $\gamma > 0$; cf. [13, sects. 1.1 and 7.1]. This implies that the objective and constraint functions of (1.1) are finite-valued for a large class of functions f_j, $j \in J$. For example, if f_j, $j \in J$, are q-times continuously differentiable, and their qth derivatives are uniformly Lipschitz continuous w.r.t. (x, ξ), we can show that the functions in (1.1) are finite-valued for all $x \in X$. We refer the reader to [20, 23, 55, 63] for further motivation to consider moment-based ambiguity sets, and to [30, 53] for discussions on the potential shortcomings of these sets.

For each lower-level optimization problem in (1.4), a worst-case distribution P_j^* is $P_j^* = N(\bar{\mu} + d_j^*, \Sigma_j^*) \in \mathcal{P}$ (see [13, sect. 7.1]), where Σ_j^* and d_j^*, respectively, are optimal solutions of (1.5) and (1.6). Generally, worst-case distributions are not unique.

We can choose the functions a_j, b_j, and C_j as $a_j = f_j(\cdot, \bar{\mu})$, $b_j = \nabla_\xi f_j(\cdot, \bar{\mu})$, and $C_j = \nabla_{\xi\xi} f_j(\cdot, \bar{\mu})$, where $\nabla_{\xi\xi} f_j(x, \bar{\mu})$ denotes the Hessian matrix of $f_j(x, \cdot)$ evaluated at $(x, \bar{\mu})$. If $x \in \mathbb{R}^n$ and the second derivative of $f_j(x, \cdot)$ is Lipschitz continuous w.r.t. ξ with Lipschitz constant $L > 0$, i.e., $|f_j(x, \xi) - m_j(x, \xi)| \leq (L/6)\|\xi - \bar{\mu}\|_2^3$, for all $\xi \in \mathbb{R}^p$, it can be shown that the worst-case expected value of the truncation error

$$\sup_{P \in \mathcal{P}} \mathbb{E}_P[|f_j(x, \xi) - m_j(x, \xi)|]$$

converges to zero as $\bar{\Sigma}_1 \to 0^+$ and $\Delta \to 0^+$. If $f_j(x, \cdot)$ are quadratic functions for each $x \in \mathbb{R}^n$ and a_j, b_j, and C_j chosen as above, the functions f_j and m_j are equal, and hence the approximation scheme is exact, i.e., (1.1) and (1.4) are equivalent.

3. Smooth DROPs, smoothing functions, and a homotopy method. We outline our algorithmic scheme to compute a stationary point of (1.4). Introducing the functions $F_j : \mathbb{R}^n \to \mathbb{R}$, $F_j(x) = \varphi_j(x) + \psi_j(x)$, $j \in J$, the DROP (1.4) becomes

$$\min_{x \in X} F_0(x) \quad \text{s.t.} \quad F_j(x) \leq 0, \quad j \in J \setminus \{0\}, \tag{3.1}$$

which is generally a nonsmooth optimization problem. In the subsequent sections, we construct smooth approximations $\widetilde{F}_j : \mathbb{R}^n \times \mathbb{R}^3_{++} \to \mathbb{R}$ of F_j parameterized by $t \in \mathbb{R}^3_{++}$. The formal definition of the functions $\widetilde{F}_j$ are given in (6.1). They are used in Algorithm 3.1 to compute a sequence of approximate KKT-points of

$$\min_{x \in X} \widetilde{F}_0(x, t) \quad \text{s.t.} \quad \widetilde{F}_j(x; t) \leq 0, \quad j \in J \setminus \{0\}, \tag{3.2}$$

as $t \to 0^+$. Since these DROPs are smooth, we can apply state-of-the-art NLP software to solve them. Throughout, let $X = \mathbb{R}^n$ hold; however, X may consist of finitely many inequality or equality constraints. Here, a point $(\bar{x}, \bar{\vartheta}) \in \mathbb{R}^n \times \mathbb{R}_+^{|J|-1}$ is referred to as a KKT-tuple of (3.1) if $\bar{\vartheta}_j F_j(\bar{x}) = 0$, $F_j(\bar{x}) \le 0$, $j \in J \setminus \{0\}$, and $0 \in \partial F_0(\bar{x}) + \sum_{j \in J\setminus\{0\}} \bar{\vartheta}_j \partial F_j(\bar{x})$. These are necessary optimality conditions for (3.1) if a constraint qualification (CQ) holds; see, e.g., [45, Cor. 5.1.8].

We construct a smoothing function of φ_j and of ψ_j that satisfies the conditions of the next definition, which is based on [18, Def. 1].

Definition 3.1. *Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. The function $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^m_{++} \to \mathbb{R}$ is referred to as a smoothing function of ϕ if $\tilde{\phi}(\cdot; t)$ is continuously differentiable for every $t > 0$, and for all $x \in \mathbb{R}^n$, it holds that*

$$\lim_{\mathbb{R}^n \ni x^k \to x,\ t^k \to 0^+} \tilde{\phi}(x^k; t^k) = \phi(x).$$

We allow for multiple smoothing parameters in Definition 3.1 as opposed to [18, Def. 1] because the smoothing function of ψ_j constructed in subsection 5.3 depends on two. In Algorithm 3.1, we do not require the computation of exact KKT-tuples of (3.2), which is important for an efficient numerical scheme for the DROP (3.1). Different notions of approximate KKT-points have been proposed in the literature; see, e.g., [1, 25]. We refer to (x, ϑ) as an ε-KKT-tuple of (3.2) if $\chi(x, \vartheta; t) \le \varepsilon$, where the criticality measure $\chi : \mathbb{R}^n \times \mathbb{R}^{|J|-1} \times \mathbb{R}^3_{++} \to \mathbb{R}_+$ is defined by

(3.3)
$$\chi(x, \vartheta; t) = \max_{j \in J\setminus\{0\}} \Big\{ \Big\| \nabla_x \widetilde{F}_0(x; t) + \sum_{j \in J\setminus\{0\}} \vartheta_j \nabla_x \widetilde{F}_j(x; t) \Big\|_\infty,\ |\min\{-\widetilde{F}_j(x; t), \vartheta_j\}| \Big\}.$$

An important notion to establish convergence of Algorithm 3.1 to stationary points of (3.1) is gradient consistency. Let $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^m_{>0} \to \mathbb{R}$ be a smoothing function of the locally Lipschitz continuous function $\phi : \mathbb{R}^n \to \mathbb{R}$. We define

$$S_{\tilde{\phi}}(x) = \operatorname{conv} \{ z \in \mathbb{R}^n :\ \exists\, \mathbb{R}^n \times \mathbb{R}^m_{++} \ni (x^k, t^k) \to (x, 0),\ \nabla_x \tilde{\phi}(x^k; t_k) \to z \}. \tag{3.4}$$

Gradient consistency of $\tilde{\phi}$ and ϕ requires the following relation to hold (cf. [14, 15, 18]):

$$S_{\tilde{\phi}}(x) = \partial \phi(x) \quad \text{for all} \quad x \in \mathbb{R}^n. \tag{3.5}$$

For the above setting, Clarke's subdifferential is a subset of (3.4) generalizing a remark in [18, sect. 1] to multiple smoothing parameters.

Lemma 3.2. *Let $\tilde{\phi} : \mathbb{R}^n \times \mathbb{R}^m_{++} \to \mathbb{R}$ be a smoothing function of the locally Lipschitz continuous function $\phi : \mathbb{R}^n \to \mathbb{R}$. Then $\partial\phi(x) \subset S_{\tilde{\phi}}(x)$ for all $x \in \mathbb{R}^n$.*

Proof. Let $x \in \mathbb{R}^n$ be arbitrary and define $\tilde{\ell} : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}$ by $\tilde{\ell}(x; t) = \tilde{\phi}(x; te)$, which is a smoothing function of ϕ, where $e = (1, \dots, 1) \in \mathbb{R}^m$. Hence, [16, Lem. 3.1] implies $\partial\phi(x) \subset S_{\tilde{\ell}}(x)$. Using (3.4), we obtain $S_{\tilde{\ell}}(x) \subset S_{\tilde{\phi}}(x)$, concluding the proof. □

In the next two sections, we construct smoothing functions of (1.5) and (1.6) that can efficiently be evaluated, as well as their gradients. Moreover, they satisfy gradient consistency.

4. Smoothing approach for the SDPs. We construct a smoothing function of φ_j (see (1.5)) satisfying the conditions stated in section 3 for the algorithmic solution of the DROP (3.1). We use the fact that the SDPs (1.5) can be solved analytically after computing the eigenvalues of a transformation of $C_j(x)$; cf. [65, Thm. 2.2].

PROPOSITION 4.1. *Let $C \in \mathbb{S}^p$ and X_0, $X_1 \in \mathbb{S}^p$ fulfill $X_0 \prec X_1$, and define $G = (X_1 - X_0)^{1/2} C (X_1 - X_0)^{1/2}$. Then it holds that*

$$C \bullet X_0 + \sum_{i=1}^{p} \min\{0, \lambda_i(G)\} = \min\{\, C \bullet X : \; X_0 \preccurlyeq X \preccurlyeq X_1 \,\}. \tag{4.1}$$

Proof. The statement follows from an application of [65, Thm. 2.2]. ∎

Numerical simulations for dimensions $p \in \{1, \dots, 2000\}$ have indicated that this solution method is significantly faster than state-of-the-art SDP solvers. If $\bar{\Sigma}_0 \prec \bar{\Sigma}_1$, then (1.5), Proposition 4.1, and (4.1) show that

$$\varphi_j(x) = (1/2) C_j(x) \bullet \bar{\Sigma}_0 + (1/2) \sum_{i=1}^{p} (\lambda_i(G_j(x)))_+ \quad \text{for all} \quad x \in \mathbb{R}^n, \tag{4.2}$$

where $G_j : \mathbb{R}^n \to \mathbb{S}^p$, $G_j(x) = (\bar{\Sigma}_1 - \bar{\Sigma}_0)^{1/2} C_j(x) (\bar{\Sigma}_1 - \bar{\Sigma}_0)^{1/2}$. In particular, φ_j is generally nonsmooth. We show that the function $\tilde{\varphi}_j : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}$ defined by

$$\tilde{\varphi}_j(x; \tau) = (1/2) C_j(x) \bullet \bar{\Sigma}_0 + (1/2) \tilde{w}(\lambda(G_j(x)); \tau) \tag{4.3}$$

is a smoothing function of φ_j, where $\tilde{w} : \mathbb{R}^n \times \mathbb{R}_{++} \to \mathbb{R}$ is given by

$$\tilde{w}(z; \tau) = \tau \sum_{i=1}^{p} \ln\left(1 + \exp\left(z_i/\tau\right)\right). \tag{4.4}$$

THEOREM 4.2. *Let $\bar{\Sigma}_0 \prec \bar{\Sigma}_1$ hold and let $C_j : \mathbb{R}^n \to \mathbb{S}^p$ be q-times continuously differentiable, where $q \geq 1$ and $j \in J$. Then the following conditions hold true:*

1. *For all $(x, \tau) \in \mathbb{R}^n \times \mathbb{R}_{++}$, we have*

$$\varphi_j(x) \leq \tilde{\varphi}_j(x; \tau) \leq \varphi_j(x) + (1/2) \tau p \ln 2, \tag{4.5}$$

   where φ_j and $\tilde{\varphi}_j$ are defined in (4.2) and (4.3), respectively.
2. *The function $\tilde{\varphi}_j$ is a smoothing function of φ_j, $\tilde{\varphi}_j(\cdot; \tau)$ is q-times continuously differentiable for every $\tau > 0$, and gradient consistency holds for $\tilde{\varphi}_j$ and φ_j.*
3. *If $(x^k) \subset \mathbb{R}^n$ and $(\tau_k) \subset \mathbb{R}_{++}$ are sequences such that $x^k \to x$ and $\tau_k \to 0$ as $k \to \infty$, there exists a convergent subsequence $(\nabla_x \tilde{\varphi}_j(x^k; \tau^k))_K$ of $(\nabla_x \tilde{\varphi}_j(x^k; \tau^k))$.*

Proof. 1. The estimate (4.5) follows from the inequalities (see, e.g., [51, sect. 2])

$$(z)_+ \leq \tau \ln(1 + \exp(z/\tau)) \leq (z)_+ + \tau \ln 2 \quad \text{for all} \quad z \in \mathbb{R}.$$

2. Next, we establish that $\tilde{\varphi}_j$ is a smoothing function of φ_j. Let $\tau > 0$ be arbitrary. The function φ_j is locally Lipschitz continuous as a composition of locally Lipschitz functions, and $\tilde{w}(\cdot; \tau)$ is symmetric and analytic as a composition of analytic functions. Hence, [60, Thm. 2.1] implies that $\tilde{w}_\lambda(\cdot; \tau) = \tilde{w}(\cdot; \tau) \circ \lambda$ is analytic, and the classical chain rule implies that $\tilde{\varphi}_j(\cdot; \tau) = (1/2) C_j(\cdot) \bullet \bar{\Sigma}_0 + (1/2) \tilde{w}_\lambda(\cdot; \tau) \circ G_j$ is q-times continuously differentiable. Together with (4.5), we obtain that $\tilde{\varphi}_j$ is a smoothing function of φ_j.

Now, we prove that gradient consistency holds, i.e., (3.5) is fulfilled. Since $\tilde{\varphi}_j$ is locally Lipschitz continuous, it suffices to show that $S_{\tilde{\varphi}_j}(x) \subset \partial \varphi_j(x)$ for all $x \in \mathbb{R}^n$; cf. Lemma 3.2, where $S_{\tilde{\varphi}_j}(x)$ is defined in (3.4). Let $x \in \mathbb{R}^n$ be arbitrary and let

$z \in \mathbb{R}^n$ be a vector such that there exist sequences $(x^k) \subset \mathbb{R}^n$ and $(\tau_k) \subset \mathbb{R}_{++}$ converging to x and 0 as $k \to \infty$, respectively, and, moreover, such that

$$\nabla_x \tilde{\varphi}_j(x^k; \tau_k) \to z \quad \text{as} \quad k \to \infty.$$

If we conclude that $z \in \partial\varphi_j(x)$, we have $S_{\tilde{\varphi}_j}(x) \subset \partial\varphi_j(x)$; see (3.4).

Now, let $k \geq 0$ be arbitrary. We compute $\nabla_x \tilde{\varphi}_j(x^k; \tau_k)$. The function $\tilde{w}(\cdot; \tau_k)$ is continuously differentiable and symmetric, and hence the classical chain rule and [40, Thm. 1.1] imply that the directional derivative $D_x\tilde{\varphi}_j(\cdot; \tau_k)h$ of $\tilde{\varphi}_j(\cdot; \tau_k)$ w.r.t. x evaluated at x^k in direction $h \in \mathbb{R}^p$ is

$$D_x\tilde{\varphi}_j(x^k; \tau_k)h = (1/2)\bar{\Sigma}_0 \bullet DC_j(x^k)h + (1/2)(Q_{j,k}M_{j,k}Q_{j,k}^T) \bullet DG_j(x^k)h,$$

where $Q_{j,k} \in \mathbb{R}^{p\times p}$ fulfills $Q_{j,k}Q_{j,k}^T = I$ and $G_j(x^k) = Q_{j,k}\mathrm{Diag}(\lambda(G_j(x^k)))Q_{j,k}^T$, and where $M_{j,k} = \mathrm{Diag}(\nabla_x\tilde{w}(\lambda(G_j(x^k)); \tau_k))$. Using the adjoint operators $DC_j(x^k)^*$ and $DG_j(x^k)^*$ of $DC_j(x^k)$ and $DG_j(x^k)$, we obtain that

$$\nabla_x\tilde{\varphi}_j(x^k; \tau_k) = (1/2)DC_j(x^k)^*\bar{\Sigma}_0 + (1/2)DG_j(x^k)^*(Q_{j,k}M_{j,k}Q_{j,k}^T). \tag{4.6}$$

We have

$$DC_j(x)^*P = \nabla_x(C_j(x) \bullet P) \quad \text{and} \quad DG_j(x^k)^*P = \nabla_x(G_j(x^k) \bullet P) \tag{4.7}$$

for all $P \in \mathbb{S}^p$. Indeed, for any $s \in \mathbb{R}^n$ and $P \in \mathbb{S}^p$, we deduce that

$$s^T DC_j(x)^*P = P \bullet DC_j(x)s = D(C_j(x) \bullet P)s = s^T\nabla_x(C_j(x) \bullet P).$$

The second equation in (4.7) can be shown similarly.

Using (4.4), we obtain

$$(\nabla_z\tilde{w}(z; \tau))_i = (1 + \exp(-z_i/\tau))^{-1} \tag{4.8}$$

for all $(z, \tau) \in \mathbb{R} \times \mathbb{R}_{++}$ and $i = 1, \ldots, p$. We deduce that $(\nabla_x\tilde{w}(\lambda(G_j(x^k)); \tau_k))$ is bounded. Moreover, $(Q_{j,k})$ is bounded. Hence, we can assume without loss of generality (w.l.o.g) that there exist $\bar{u}^j \in \mathbb{R}^p$ and $\bar{Q}_j \in \mathbb{R}^{p\times p}$ such that

$$\nabla_x\tilde{w}(\lambda(G_j(x^k)); \tau_k) \to \bar{u}^j \quad \text{and} \quad Q_{j,k} \to \bar{Q}_j \quad \text{as} \quad k \to \infty,$$

with $\bar{Q}_j\bar{Q}_j^T = I$ and $G_j(x) = \bar{Q}_j\mathrm{Diag}(\lambda(G_j(x)))\bar{Q}_j^T$, where we have used that λ is continuous; cf. [33, Cor. 6.3.8]. In addition, (4.8) implies for $i = 1, \ldots, p$ that

$$(\nabla_x\tilde{w}(\lambda(G_j(x^k)); \tau_k))_i \to (\bar{u}^j)_i \in \begin{cases} \{0\} & \text{if} \quad \lambda_i(G_j(x)) < 0, \\ [0, 1] & \text{if} \quad \lambda_i(G_j(x)) = 0, \\ \{1\} & \text{if} \quad \lambda_i(G_j(x)) > 0, \end{cases} \quad \text{as} \quad k \to \infty.$$

Hence, (4.6) and the continuity of both DC_j and DG_j show that

$$\nabla_x\tilde{\varphi}_j(x^k; \tau_k) \to (1/2)DC_j(x)^*\bar{\Sigma}_0 + (1/2)DG_j(x)^*\bar{Q}_j\mathrm{Diag}(\bar{u}^j)\bar{Q}_j^T = z \quad \text{as} \quad k \to \infty.$$

To verify that $z \in \partial\varphi_j(x)$, we compute $\partial\varphi_j(x)$ using (4.2). The function $\mathbb{S}^p \ni G \mapsto \sum_{i=1}^p (\lambda_i(G))_+$ is regular (cf. [41, Cor. 4]), sums of regular functions are regular, and continuously differentiable functions are regular; cf. [22, Prop. 2.3.6]. Hence, through applications of the chain rule [21, Thm. 2.3.10] and [41, Thm. 8], we obtain that

$$\partial\varphi_j(x) = \left\{ \tfrac{1}{2}DC_j(x)^*\bar{\Sigma}_0 + \tfrac{1}{2}DG_j(x)^*Q\mathrm{Diag}(u)Q^T : Q \in O_j(x), u \in \partial w(\lambda(G_j(x))) \right\}, \tag{4.9}$$

where $O_j(x) = \{Q \in \mathbb{R}^{p\times p} : QQ^T = I, G_j(x) = Q\mathrm{Diag}(\lambda(G_j(x)))Q^T\}$ and $w : \mathbb{R}^p \to \mathbb{R}$ is defined by $w(z) = \sum_{i=1}^p (z)_+$. For each $z \in \mathbb{R}^p$, and for all $i \in \{1,\dots,p\}$ and $g \in \partial w(z)$, it holds that $g_i = 0$ if $z_i < 0$, $g_i \in [0,1]$ if $z_i = 0$, and $g_i = 1$ if $z_i > 0$. Hence, we deduce $\bar{u}^j \in \partial w(\lambda(G_j(x)))$ and, finally, that $z \in \partial\varphi_j(x)$.

3. We can adapt the above reasoning to deduce that $(\nabla_x\tilde{\varphi}_j(x^k;\tau^k))$ has a convergent subsequence if $(x^k) \subset \mathbb{R}^n$ and $(\tau_k) \subset \mathbb{R}_{++}$ fulfill $x^k \to x$, $\tau_k \to 0$ as $k \to \infty$. ∎

Based on an eigendecomposition of $G_j(x)$, the computation of $\nabla_x\tilde{\varphi}_j(x;\tau)$ is cheap; cf. (4.6). The next step in order to solve the DROP (3.1) efficiently is to construct a computationally tractable smoothing function of (1.6).

5. Smoothing approach for the TRPs.

We derive a smoothing function of the optimal value function defined in (1.6) by constructing one for the function $v : \mathbb{R}^n \to \mathbb{R}$ defined by

$$v(x) = \min_{s\in\mathbb{R}^p} \left\{ (1/2)s^T H(x)s + g(x)^T s : \quad (1/2)\|s\|_2^2 \le (1/2)\Delta^2 \right\}, \tag{5.1}$$

where $g : \mathbb{R}^n \to \mathbb{R}^p$ and $H : \mathbb{R}^n \to \mathbb{S}^p$. Throughout, let $\Delta > 0$ be satisfied. We obtain a smoothing function of (5.1) as a value function of a "lifted" TRP. The lifted TRP results from a barrier formulation of a Lagrangian dual of (5.1). Since TRPs are theoretically and practically tractable (see [6, sect. 2] and [47, sect. 5]), our construction implies that the smoothing function of v can be evaluated efficiently. Moreover, based on Danskin's theorem, we can deduce that the evaluations of derivatives of the smoothing function are computationally tractable as well. In addition, we establish gradient consistency, and thus the smoothing function meets the conditions stated in section 3. In particular, we deduce that the DROP (3.1) can be solved by Algorithm 3.1. Our approximation and smoothing scheme can be applied to nonlinear ROPs as an alternative to methods used in, e.g., [24, 39].

5.1. Lagrangian dual of TRPs.

Before we review properties of the Lagrangian dual of the nominal TRP

$$\min_{s\in\mathbb{R}^p} (1/2)s^T Hs + g^T s \quad \text{s.t.} \quad (1/2)\|s\|_2^2 \le (1/2)\Delta^2, \tag{5.2}$$

where $g = g(x_0) \in \mathbb{R}^p$, $H = H(x_0) \in \mathbb{S}^p$, and $x_0 \in \mathbb{R}^n$, we state necessary and sufficient optimality conditions of (5.2); see, e.g., [56, Lems. 2.4 and 2.8].

THEOREM 5.1. *The TRP* (5.2) *has an optimal solution* $s^* \in \mathbb{R}^p$. *Moreover, the vector* $s^* \in \mathbb{R}^p$ *is an optimal solution of* (5.2) *iff there exists* $\lambda^* \in \mathbb{R}$ *such that*

$$(H + \lambda^* I)s^* = -g, \quad \|s^*\|_2 \le \Delta, \quad \lambda^*(\|s^*\|_2 - \Delta) = 0, \quad \lambda^* \ge 0, \quad H + \lambda^* I \succcurlyeq 0. \tag{5.3}$$

In addition, if (s^*, λ^*) *fulfills* (5.3) *and* $\lambda^* > -\lambda_{\min}(H)$, *then* s^* *is the unique optimal solution of* (5.2). *Moreover, if* (s_1^*, λ_1^*) *and* (s_2^*, λ_2^*) *fulfill* (5.3), *it holds that* $\lambda_1^* = \lambda_2^*$.

If (s^*, λ^*) satisfies (5.3), we refer to it as the optimal primal-dual solution of (5.2). Next, we provide a definition of the hard case of the TRP (5.2).

DEFINITION 5.2. *Let* (s^*, λ^*) *be an optimal primal-dual solution of* (5.2). *If* $\lambda^* = -\lambda_{\min}(H)$ *holds, the* hard case *occurs for* (5.2), *and otherwise the* easy case *occurs.*

The term "hard case" is due to [47], and the terminology of "easy case" has been used in, e.g., [58]. Now, we state a result on Lagrangian duality of (5.2).

THEOREM 5.3 (see [59, Prop. 3.1, Thm. 3.3, Cor. 3.4]). *A Lagrangian dual problem of* (5.2)*—phrased as a minimization problem—is given by*

$$\min_{\lambda\in\mathbb{R}} d(\lambda) \quad s.t. \quad H+\lambda I \succcurlyeq 0, \quad \lambda \geq 0, \tag{5.4}$$

where $d : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ *is defined by*

$$d(\lambda) = \begin{cases} \frac{1}{2} g^T (H+\lambda I)^+ g + \frac{1}{2}\Delta^2 \lambda & \textit{if} \quad \lambda \geq (-\lambda_{\min}(H))_+,\ g \perp N(H+\lambda I), \\ \infty & \textit{else.} \end{cases} \tag{5.5}$$

Moreover, (5.4) *has a unique optimal solution* λ^*, *which is the unique Lagrange multiplier associated to* (5.2). *In addition, strong duality holds, i.e., the optimal value of* (5.2) *equals* $-d^*$, *where* d^* *denotes the optimal value of* (5.4).

We define the solution mapping $s : \mathbb{R} \to \mathbb{R}^p$ by

$$s(\lambda) = -(H+\lambda I)^+ g \tag{5.6}$$

and summarize properties of the dual function d.

LEMMA 5.4. *The following conditions hold true:*

1. *The function* d *defined in* (5.5) *is convex and* $d(\lambda) \to \infty$ *as* $\lambda \to \infty$.
2. *If* $\lambda > (-\lambda_{\min}(H))_+$, *then* d *is twice continuously differentiable at* λ, *and*
$$d'(\lambda) = -(1/2)\|s(\lambda)\|_2^2 + (1/2)\Delta^2. \tag{5.7}$$
3. *If* $g \neq 0$, *then* $d''(\lambda) > 0$ *for all* $\lambda > (-\lambda_{\min}(H))_+$.

Proof. The statements follow from [59, Prop. 3.2] and from the proof of [59, Thm. 3.3]. ∎

5.2. Barrier formulation for the dual of TRPs. We state a barrier problem of (5.4) using a reciprocal barrier and show that an optimal solution of it is an approximate solution to (5.4). In subsection 5.3, it is shown that the barrier problem corresponds to a "lifted" TRP which justifies the use of a reciprocal barrier instead of a self-concordant one. Hence, it can be solved with any TRP solver, enabling us to define and evaluate a smoothing function of ψ_j (see (1.6)) and its derivatives efficiently and, subsequently, to solve the DROP (3.1). The barrier problem associated to (5.4) is

$$\min_{\lambda\in\mathbb{R}} d(\lambda) + \nu B_\eta(\lambda) \quad \text{s.t.} \quad \lambda > E(-H;\eta), \quad \lambda > 0, \tag{5.8}$$

where ν, $\eta > 0$ and the reciprocal barrier $B_\eta : ((E(-H;\eta))_+, \infty) \to \mathbb{R}$ is defined by

$$B_\eta(\lambda) = \frac{1}{\lambda} + \frac{1}{\lambda - E(-H;\eta)}; \tag{5.9}$$

see, e.g., [29, sect. 3.1]. Here, $E : \mathbb{S}^p \times \mathbb{R}_{++} \to \mathbb{R}$ is an entropy function defined by

$$E(A;\eta) = \eta \ln \sum_{i=1}^{p} \exp(\lambda_i(A)/\eta). \tag{5.10}$$

It has been successfully used in the context of nonsmooth optimization (see, e.g., [19, 48]); E is a smoothing function of $\lambda_{\max}$ and fulfills

$$\lambda_{\max}(A) \le E(A;\eta) \le \lambda_{\max}(A) + \eta \ln p \tag{5.11}$$

for all $A \in \mathbb{S}^p$ and every $\eta > 0$; cf. [48, eqns. (17) and (18)] and [33, Cor. 6.3.8]. In particular, for all $A \in \mathbb{S}^p$ and any $\eta > 0$, we have

$$\lambda_{\min}(A) = -\lambda_{\max}(-A) \ge -E(-A;\eta). \tag{5.12}$$

We could use the barrier function $((-\lambda_{\min}(H))_+, \infty) \ni \lambda \mapsto -\ln\lambda - \ln\det(H+\lambda I)$ in (5.8), which does not require the computation of $\lambda_{\min}(H)$ and to smooth $\lambda_{\min}$. However, the resulting primal problem would not be a TRP and requires, e.g., an adapted version of [47, Alg. 3.2] for its numerical solution. Next, we show that (5.8) has a unique optimal solution for any ν, $\eta > 0$.

LEMMA 5.5. *For every ν, $\eta > 0$, the barrier problem* (5.8) *has a unique optimal solution $\lambda^*(\nu,\eta)$, and $\lambda^*(\nu,\eta) > (E(-H;\eta))_+$ holds, where E is defined in* (5.10).

Proof. Let ν, $\eta > 0$ be arbitrary. Define the objective function of (5.8) by

$$B_{\nu,\eta} : ((E(-H;\eta))_+, \infty) \to \mathbb{R}, \quad B_{\nu,\eta} = d + \nu B_\eta, \tag{5.13}$$

where d and B_η are defined in (5.5) and (5.9), respectively. Let $\lambda > (E(-H;\eta))_+$ be arbitrary. Since $(E(-H;\eta))_+ \ge (-\lambda_{\min}(H))_+$ holds (cf. (5.12)), we have

$$B_{\nu,\eta}(\lambda) = \frac{1}{2}g^T(H+\lambda I)^{-1}g + \frac{1}{2}\Delta^2\lambda + \frac{\nu}{\lambda} + \frac{\nu}{\lambda - E(-H;\eta)} \ge \frac{1}{2}\Delta^2\lambda,$$

showing that $B_{\nu,\eta}(\lambda) \to \infty$ as $\lambda \to \infty$. From (5.5), (5.9), and (5.13), we deduce that

$$B_{\nu,\eta}(\lambda) \ge \frac{\nu}{\lambda} + \frac{\nu}{\lambda - E(-H;\eta)} \to \infty \quad \text{as} \quad \lambda \to (E(-H;\eta))_+.$$

Thus, (5.8) has an optimal solution $\lambda^*(\nu,\eta)$, and $\lambda^*(\nu,\eta) > (E(-H;\eta))_+$ holds.

Now, we show that $B_{\nu,\eta}$ is strictly convex. Lemma 5.4 implies that $B_{\nu,\eta}$ (cf. (5.13)) is twice continuously differentiable at λ with

$$B'_{\nu,\eta}(\lambda) = -\frac{1}{2}g^T(H+\lambda I)^{-2}g - \frac{\nu}{\lambda^2} - \frac{\nu}{(\lambda - E(-H;\eta))^2} + \frac{1}{2}\Delta^2 \tag{5.14}$$

and

$$B''_{\nu,\eta}(\lambda) = g^T(H+\lambda I)^{-3}g + \frac{2\nu}{\lambda^3} + \frac{2\nu}{(\lambda - E(-H;\eta))^3} > 0,$$

implying that $B_{\nu,\eta}$ is strictly convex. Hence, $\lambda^*(\nu,\eta)$ is the unique solution of (5.8). ◻

For ν, $\eta > 0$, we denote by $\lambda^*(\nu,\eta)$ the optimal solution of (5.8); cf. Lemma 5.5.

THEOREM 5.6. *Let ν, $\eta > 0$ be arbitrary. Then the following conditions hold:*

1. *We have*

$$\lambda^*(\nu,\eta) \ge \sqrt{2\nu}/\Delta \quad \textit{and} \quad \lambda^*(\nu,\eta) - E(-H;\eta) \ge \sqrt{2\nu}/\Delta, \tag{5.15}$$

where $\lambda^*(\nu,\eta)$ is the optimal solution of (5.8) *and E is defined in* (5.10).

2. *The point* $\lambda^*(\nu,\eta)$ *is a* $(\sqrt{2\nu}\Delta + (1/2)\Delta^2\eta\ln p)$*-optimal solution of* (5.4), *i.e.,*

$$d^* \le d(\lambda^*(\nu,\eta)) \le d^* + \sqrt{2\nu}\Delta + (1/2)\Delta^2\eta\ln p, \tag{5.16}$$

where d^* *denotes the optimal value of* (5.4) *and* d *is defined in* (5.5).

3. *It holds that*

$$d^* \le d(\lambda^*(\nu,\eta)) + \nu B_\eta(\lambda^*(\nu,\eta)) \le d^* + 2\sqrt{2\nu}\Delta + (1/2)\Delta^2\eta\ln p, \tag{5.17}$$

where the barrier function B_η *is defined in* (5.9).

We apply the following result to prove Theorem 5.6.

LEMMA 5.7. *Let* η, $\epsilon > 0$ *be arbitrary, and consider*

$$\min_{\lambda\in\mathbb{R}} d(\lambda) \quad s.t. \quad \lambda \ge \epsilon, \quad \lambda \ge E(-H;\eta) + \epsilon. \tag{5.18}$$

Then problem (5.18) *has a unique optimal solution* $\bar\lambda_{\eta,\epsilon}$. *Moreover, it holds that*

$$d^* \le d(\bar\lambda_{\eta,\epsilon}) = d^*_{\eta,\epsilon} \le d^* + (1/2)\Delta^2(\eta\ln p + \epsilon), \tag{5.19}$$

where d^* *denotes the optimal value of* (5.4) *and* $d^*_{\eta,\epsilon}$ *that of* (5.18).

Proof. We establish existence and uniqueness of solutions of (5.18). If $g = 0$, we obtain $d(\lambda) = (1/2)\Delta^2\lambda$. Hence, the optimal solution $\bar\lambda_{\eta,\epsilon}$ of (5.18) is given by $\bar\lambda_{\eta,\epsilon} = (E(-H;\eta))_+ + \epsilon$. If $g \ne 0$, Lemma 5.4 and (5.12) imply that the objective of (5.18) is coercive, twice continuously differentiable in an open neighborhood of the feasible set of (5.18), and $d''(\lambda) > 0$ for all $\lambda > (E(-H;\eta))_+$. Hence, there exists a unique optimal solution $\bar\lambda_{\eta,\epsilon}$ of (5.18).

Now, we establish (5.19). Since $\bar\lambda_{\eta,\epsilon} \ge (E(-H;\eta))_+ + \epsilon$, we have $d^* \le d(\bar\lambda_{\eta,\epsilon})$. Moreover, if $\lambda^* > (E(-H;\eta))_+ + \epsilon$ holds, we deduce $d^* = d^*_{\eta,\epsilon}$, where λ^* denotes the optimal solution of (5.4). Hence, the remaining case to be considered is

$$(-\lambda_{\min}(H))_+ \le \lambda^* \le (E(-H;\eta))_+ + \epsilon.$$

We define $\bar\lambda = \lambda^* + \eta\ln p + \epsilon$ and observe that $\bar\lambda \ge \epsilon$. From (5.11), we deduce that

$$E(-H;\eta) \le -\lambda_{\min}(H) + \eta\ln p \le \lambda^* + \eta\ln p,$$

showing that $\bar\lambda \ge E(-H;\eta) + \epsilon$. Hence, $\bar\lambda$ is feasible for (5.18). Lemma 5.4 implies that d is convex, and differentiable at $\bar\lambda$. Therefore, we have

$$d(\lambda^*) - d(\bar\lambda) \ge d'(\bar\lambda)(\lambda^* - \bar\lambda) = -d'(\bar\lambda)(\eta\ln p + \epsilon),$$

resulting in

$$d(\lambda^*) + d'(\bar\lambda)(\eta\ln p + \epsilon) \ge d(\bar\lambda) \ge d(\bar\lambda_{\eta,\epsilon}).$$

Now, (5.6), Lemma 5.4, and (5.7) imply $d'(\bar\lambda) \le (1/2)\Delta^2$, and hence (5.19) holds. ∎

To prove the estimates in (5.16), we use the fact that the functions $G_1 : (0,\infty) \to \mathbb{R}$ and $G_2 : (E(-H;\eta),\infty) \to \mathbb{R}$ defined by

$$G_1(\lambda) = -\ln\lambda \quad \text{and} \quad G_2(\lambda) = -\ln(\lambda - E(-H;\eta))$$

are 1-self-concordant barrier functions of their domains; cf. [49, sect. 2.3.1, Ex. 2].

Proof of Theorem 5.6. 1. We establish (5.15). Recall that the objective of (5.8) is $B_{\nu,\eta}$; cf. (5.13). Lemma 5.5 implies that $B'_{\nu,\eta}(\lambda^*(\nu,\eta)) = 0$ and (5.14) results in

$$g^T(H + \lambda^*(\nu,\eta)I)^{-2}g + \frac{2\nu}{\lambda^*(\nu,\eta)^2} + \frac{2\nu}{(\lambda^*(\nu,\eta) - E(-H;\eta))^2} = \Delta^2.$$

Lemma 5.5 and (5.12) further yield $H + \lambda^*(\nu,\eta)I \in \mathbb{S}^p_{++}$, and hence we deduce that

$$\frac{2\nu}{\lambda^*(\nu,\eta)^2} \le \Delta^2 \quad \text{and} \quad \frac{2\nu}{(\lambda^*(\nu,\eta) - E(-H;\eta))^2} \le \Delta^2,$$

showing the estimates in (5.15).

2. Next, we verify (5.16). The point $\lambda^*(\nu,\eta)$ is feasible for (5.4) by (5.15), and therefore we have $d^* \le d(\lambda^*(\nu,\eta))$. Now, let $\lambda > (E(-H;\eta))_+$ be arbitrary. Both functions G_1 and G_2 defined prior the proof are 1-self-concordant for their domains. Hence, we obtain from [49, Prop. 2.3.2] that

$$\begin{aligned} -\frac{1}{\lambda^*(\nu,\eta)}(\lambda - \lambda^*(\nu,\eta)) &= G_1'(\lambda^*(\nu,\eta))(\lambda - \lambda^*(\nu,\eta)) \le 1, \\ -\frac{1}{\lambda^*(\nu,\eta) - E(-H;\eta)}(\lambda - \lambda^*(\nu,\eta)) &= G_2'(\lambda^*(\nu,\eta))(\lambda - \lambda^*(\nu,\eta)) \le 1. \end{aligned} \tag{5.20}$$

Further, $B'_{\nu,\eta}(\lambda^*(\nu,\eta)) = 0$ results in

$$d'(\lambda^*(\nu,\eta)) = -\nu B'_\eta(\lambda^*(\nu,\eta)),$$

showing with (5.15), (5.20), and $\lambda^*(\nu,\eta) > (E(-H;\eta))_+$ that

$$\begin{aligned} d'(\lambda^*(\nu,\eta))(\lambda - \lambda^*(\nu,\eta)) &= -\nu B'_\eta(\lambda^*(\nu,\eta))(\lambda - \lambda^*(\nu,\eta)) \\ &= \frac{\nu}{\lambda^*(\nu,\eta)^2}(\lambda - \lambda^*(\nu,\eta)) + \frac{\nu}{(\lambda^*(\nu,\eta) - E(-H;\eta))^2}(\lambda - \lambda^*(\nu,\eta)) \\ &\ge -\frac{\nu}{\lambda^*(\nu,\eta)} - \frac{\nu}{\lambda^*(\nu,\eta) - E(-H;\eta)}. \end{aligned}$$

Next, the convexity of d (cf. Lemma 5.4), the above formula, and (5.15) yield that

$$\begin{aligned} d(\lambda^*(\nu,\eta)) - d(\lambda) &\le d'(\lambda^*(\nu,\eta))(\lambda^*(\nu,\eta) - \lambda) \\ &\le \frac{\nu}{\lambda^*(\nu,\eta)} + \frac{\nu}{\lambda^*(\nu,\eta) - E(-H;\eta)} \le \frac{2\nu}{\sqrt{2\nu}}\Delta = \sqrt{2\nu}\Delta. \end{aligned} \tag{5.21}$$

Now, we denote by $\bar{\lambda}_{\eta,\epsilon}$ the optimal solution of (5.18) for an arbitrary $\epsilon > 0$, which fulfills $\bar{\lambda}_{\eta,\epsilon} \ge (E(-H;\eta))_+ + \epsilon$; cf. Lemma 5.7. Furthermore, Lemma 5.7, (5.19), and (5.21) with $\lambda = \bar{\lambda}_{\eta,\epsilon}$ show that

$$d(\lambda^*(\nu,\eta)) \le d(\bar{\lambda}_{\eta,\epsilon}) + \sqrt{2\nu}\Delta \le d^* + \sqrt{2\nu}\Delta + (1/2)\Delta^2(\eta \ln p + \epsilon).$$

The latter inequalities hold for all $\epsilon > 0$, and hence we obtain (5.16).

3. We show (5.17). Using (5.9) and (5.15), we deduce that $\nu B_\eta(\lambda^*(\nu,\eta)) > 0$, $\nu B_\eta(\lambda^*(\nu,\eta)) \le \sqrt{2\nu}\Delta$, and $\lambda^*(\nu,\eta)$ is feasible for (5.4). Hence, (5.16) implies (5.17). ∎

The error estimates presented in Theorem 5.6 depend on $\ln p$ and on the prescribed trust-region radius Δ. Therefore, the data dependence is weak.

5.3. Smoothing function for TRPs. We show that the function $\tilde{v} : \mathbb{R}^n \times \mathbb{R}^2_{++} \to \mathbb{R}$ defined by

$$\tilde{v}(x;\nu,\eta) = \min_{\widetilde{s}\in\mathbb{R}^{p+2}} \left\{ (1/2)\widetilde{s}^T \widetilde{H}_\eta(x)\widetilde{s} + \widetilde{g}_\nu(x)^T\widetilde{s} : \ \ (1/2)\|\widetilde{s}\|_2^2 \le (1/2)\Delta^2 \right\} \tag{5.22}$$

is a smoothing function of v (see (5.1)) and establish gradient consistency, where

$$\widetilde{H}_\eta(x) = \begin{bmatrix} H(x) & & \\ & 0 & \\ & & -E(-H(x);\eta) \end{bmatrix} \in \mathbb{S}^{p+2} \ \text{ and } \ \widetilde{g}_\nu(x) = \begin{bmatrix} g(x) \\ \sqrt{2\nu} \\ \sqrt{2\nu} \end{bmatrix} \in \mathbb{R}^{p+2}, \tag{5.23}$$

and $E(\cdot;\eta)$ is defined in (5.10). Subsequently, we apply these results to define a smoothing function of ψ_j (see (1.6)), to deduce its gradient consistency, and to deduce computationally tractability—crucial properties for an efficient solution of approximated DROPs using Algorithm 3.1. To prove these properties, we use the fact that a Lagrangian dual of (5.22) is

$$\min_{\lambda\in\mathbb{R}} d(\lambda;x) + \frac{\nu}{\lambda} + \frac{\nu}{\lambda - E(-H(x);\eta)} \quad \text{s.t.} \quad \lambda > E(-H(x);\eta), \quad \lambda > 0, \tag{5.24}$$

where $x \in \mathbb{R}^n$ and $d : ((-E(H(x);\eta))_+, \infty) \times \mathbb{R}^n \to \mathbb{R}$ is defined by

$$d(\lambda;x) = (1/2)g(x)^T(H(x)+\lambda I)^{-1}g(x) + (1/2)\Delta^2\lambda. \tag{5.25}$$

LEMMA 5.8. *Let $x \in \mathbb{R}^n$ and ν, $\eta > 0$ be arbitrary. Then problem* (5.24) *has a unique optimal solution $\tilde{\lambda}(x;\nu,\eta)$, and it holds that $\tilde{\lambda}(x;\nu,\eta) > (E(-H(x);\eta))_+$. Moreover, the optimal value of* (5.22) *equals the negative of that of* (5.24), *the hard case does not occur for* (5.22), *and*

$$\tilde{v}(x;\nu,\eta) = -(1/2)\widetilde{g}_\nu(x)^T(\widetilde{H}_\eta(x) + \tilde{\lambda}(x;\nu,\eta)I)^{-1}\widetilde{g}_\nu(x) - (1/2)\Delta^2\tilde{\lambda}(x;\nu,\eta). \tag{5.26}$$

Proof. Lemma 5.5 implies that (5.24) has a unique optimal solution $\tilde{\lambda}(x;\nu,\eta)$, and it holds that $\tilde{\lambda}(x;\nu,\eta) > (E(-H(x);\eta))_+$. Using (5.12), we deduce that $\lambda_{\min}(H(x)) \ge -E(-H(x);\eta)$ and (5.23) shows $\lambda_{\min}(\widetilde{H}_\eta(x)) = -(E(-H(x);\eta))_+$.

If $E(-H(x);\eta) > 0$, we have $y = (0,\dots,0,1) \in N(\widetilde{H}_\eta(x) - \lambda_{\min}(\widetilde{H}_\eta(x))I)$ and $y^T\widetilde{g}_\nu(x) \neq 0$. If $E(-H(x);\eta) \le 0$, we get $w = (0,\dots,0,1,0) \in N(\widetilde{H}_\eta(x) - \lambda_{\min}(\widetilde{H}_\eta(x))I)$ and $w^T\widetilde{g}_\nu(x) \neq 0$. Hence, we obtain $\widetilde{g}_\nu(x) \not\perp N(\widetilde{H}_\eta(x) - \lambda_{\min}(\widetilde{H}_\eta(x))I)$.

Next, for all $\lambda > (E(-H(x);\eta))_+$, we deduce from (5.23) and (5.25) that

$$d(\lambda;x) + \frac{\nu}{\lambda(x;\nu,\eta)} + \frac{\nu}{\lambda(x;\nu,\eta) - E(-H(x);\eta)} = \frac{1}{2}\widetilde{g}_\nu(x)^T(\widetilde{H}_\eta(x) + \lambda I)^{-1}\widetilde{g}_\nu(x) + \frac{1}{2}\Delta^2\lambda.$$

Hence, Theorem 5.3 shows that strong duality holds and (5.26) is satisfied. The hard case does not occur for (5.22) since $\tilde{\lambda}(x;\nu,\eta) > (E(-H(x);\eta))_+ = -\lambda_{\min}(\widetilde{H}_\eta(x))$. $\square$

We establish an error estimate on $\tilde{v}$ (see (5.22)) and show that it is a smoothing function of v (see (5.1)). We define, similarly to (5.6), the mapping $s : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ by

$$s(\lambda;x) = -(H(x) + \lambda I)^+ g(x). \tag{5.27}$$

For ν, $\eta > 0$, we denote by $(\widetilde{s}(x;\nu,\eta), \tilde{\lambda}(x;\nu,\eta))$ an optimal primal-dual solution of (5.22), where

$$\tilde{\lambda}(\cdot;\nu,\eta) : \mathbb{R}^n \to \mathbb{R} \quad \text{and} \quad \widetilde{s}(\cdot;\nu,\eta) : \mathbb{R}^n \to \mathbb{R}^p. \tag{5.28}$$

From (5.3), Lemma 5.8, the block structure of $\widetilde{H}_\eta(x)$ (see (5.23)), and (5.27), we deduce that for all $x \in \mathbb{R}^n$ it holds that

$$\widetilde{s}(x;\nu,\eta) = (s(\tilde{\lambda}(x;\nu,\eta);x), \widetilde{s}_{p+1}(x;\nu,\eta), \widetilde{s}_{p+2}(x;\nu,\eta)). \tag{5.29}$$

In particular, the first p components of $\widetilde{s}(x;\nu,\eta)$ are given by $s(\tilde{\lambda}(x;\nu,\eta);x)$. By applying (5.3) and (5.23), we obtain that

$$\widetilde{s}_{p+1}(x;\nu,\eta) = \frac{\sqrt{2\nu}}{\tilde{\lambda}(x;\nu,\eta)} \quad \text{and} \quad \widetilde{s}_{p+2}(x;\nu,\eta) = \frac{\sqrt{2\nu}}{\tilde{\lambda}(x;\nu,\eta) - E(-H(x);\eta)}. \tag{5.30}$$

THEOREM 5.9. *Let ν, $\eta > 0$ be arbitrary, and let the mappings $g : \mathbb{R}^n \to \mathbb{R}^p$ and $H : \mathbb{R}^n \to \mathbb{S}^p$ be q-times continuously differentiable, where $q \geq 1$. Then the following conditions hold true:*

1. *For every $x \in \mathbb{R}^n$, we have*

$$v(x) \geq \tilde{v}(x;\nu,\eta) \geq v(x) - 2\sqrt{2\nu}\Delta - (1/2)\Delta^2\eta \ln p, \tag{5.31}$$

where v is defined in (5.1) *and $\tilde{v}$ in* (5.22).

2. *The mappings $\widetilde{s}(\cdot;\nu,\eta)$ and $\tilde{\lambda}(\cdot;\nu,\eta)$ defined in* (5.28) *are $q-1$-times continuously differentiable, and $\tilde{v}(\cdot;\nu,\eta)$ is q-times continuously differentiable. We have*

$$\nabla_x \tilde{v}(x;\nu,\eta) = \nabla_x \wp(x,s)|_{s=s(\tilde{\lambda};x)} + (1/2)(\widetilde{s}_{p+2})^2 \nabla_x(-E(-H(x);\eta)), \tag{5.32}$$

where $\wp : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}$ is defined by

$$\wp(x,s) = g(x)^T s + (1/2)s^T H(x) s \tag{5.33}$$

and $(\widetilde{s}, \tilde{\lambda}) = (\widetilde{s}(x;\nu,\eta), \tilde{\lambda}(x;\nu,\eta))$ is the optimal primal-dual solution of (5.22).

3. *The function $\tilde{v}$ is a smoothing function of v.*

Proof. 1. Let $x \in \mathbb{R}^n$ be arbitrary. Theorem 5.6 and Lemma 5.8 yield with (5.17) and (5.26) that (5.31) holds.

2. Lemma 5.8 further shows that $\tilde{\lambda}(x;\nu,\eta) > (E(-H(x);\eta))_+$, implying that strict complementarity slackness holds for (5.22). Moreover, the function $E(\cdot;\eta)$ (see (5.10)) is analytic as $z \mapsto \eta \ln \sum_{i=1}^p \exp(z_i/\eta)$ is analytic (see [60, Thm. 3.1]), and therefore the mapping $\widetilde{H}_\eta$ (see (5.23)) is q-times continuously differentiable. Hence, the implicit function theorem applies to the first order optimality conditions (5.3) of (5.22) and implies that $\tilde{\lambda}(\cdot;\nu,\eta)$ and $\widetilde{s}(\cdot;\nu,\eta)$ are $q-1$-times continuously differentiable.

Now, (5.22), (5.23), (5.29), and (5.33), together with Danskin's theorem [11, Thm. 4.13, Rem. 4.14], yield that $\tilde{v}(\cdot;\nu,\eta)$ is differentiable and show that its gradient is given by (5.32). Next, [32, Cor. 8.2] implies that $\widetilde{s}(\cdot;\nu,\eta)$ is continuous, showing that $\nabla_x \tilde{v}(\cdot;\nu,\eta)$ is continuous. Moreover, the chain rule and (5.22) imply that $\tilde{v}(\cdot;\nu,\eta)$ is q-times continuously differentiable.

3. The function v is continuous by [32, Thm. 7], $\tilde{v}(\cdot;\nu,\eta)$ is continuously differentiable, and hence (5.31) shows that $\tilde{v}$ is a smoothing function of v. ∎

The next result asserts gradient consistency of the function $\tilde{v}$ defined in (5.22).

THEOREM 5.10. *Let the conditions of Theorem 5.9 be fulfilled. Then the following conditions are satisfied:*

1. *Gradient consistency holds for $\tilde{v}$ and v, where v is defined in (5.1) and $\tilde{v}$ in (5.22).*
2. *Let $x \in \mathbb{R}^n$ be given, let $(x^k) \subset \mathbb{R}^n$, and let (ν_k), $(\eta_k) \subset \mathbb{R}_{++}$ be sequences converging to x and 0 as $k \to \infty$, respectively. Then there exists a convergent subsequence $(\nabla_x \tilde{v}(x^k; \nu_k, \eta_k))_K$ of $(\nabla_x \tilde{v}(x^k; \nu_k, \eta_k))$.*

We need the following result to prove Theorem 5.10.

LEMMA 5.11. *Let $(\eta_k) \subset \mathbb{R}_{++}$ be a sequence such that $\eta_k \to 0$ as $k \to \infty$. Furthermore, let $A : \mathbb{R}^n \to \mathbb{S}^p$ be continuously differentiable and let $(x^k) \subset \mathbb{R}^n$ be a sequence such that $x^k \to x \in \mathbb{R}^n$ as $k \to \infty$. Then there exist a subsequence $(\nabla_x(E(\cdot;\eta_k) \circ A)(x^k))_K$ of $(\nabla_x(E(\cdot;\eta_k) \circ A)(x^k))$, $\theta_i \in [0,1]$, and $u_i \in \mathbb{R}^p$, such that*

$$\nabla_x(E(\cdot;\eta_k) \circ A)(x^k) \to \sum_{i=1}^{r} \theta_i DA(x)^*[u_i u_i^T] \in DA(x)^* \partial \lambda_{\max}(A(x)) \text{ as } K \ni k \to \infty,$$

where E is defined in (5.10), $1 \le r \le r(A(x))$, $r(A(x))$ denotes the multiplicity of $\lambda_{\max}(A(x))$, $\sum_{i=1}^{r} \theta_i = 1$, and $\|u_i\|_2 = 1$ are pairwise orthogonal eigenvectors of $A(x)$ corresponding to $\lambda_{\max}(A(x))$.

Proof. The mapping A is continuously differentiable and $\lambda_{\max}$ is convex and, hence, regular in the sense of [22, Def. 2.3.4]; see [22, Prop. 2.3.6]. Moreover, A and $\lambda_{\max}$ are locally Lipschitz continuous. The chain rule [22, Thm. 2.3.9] implies that

$$\partial(\lambda_{\max} \circ A)(x) = DA(x)^* \partial \lambda_{\max}(A(x)). \tag{5.34}$$

The function $E(\cdot;\eta_k)$ is analytic (see [60, Thm. 3.1]), and hence the chain rule implies

$$\nabla_x(E(\cdot;\eta_k) \circ A)(x^k) = DA(x^k)^* \nabla_A E(A(x^k);\eta_k). \tag{5.35}$$

We define $A_k = A(x^k)$ and $A = A(x)$. Next, we show that there exists a subsequence $(\nabla_A E(A_k;\eta_k))_K$ of $(\nabla_A E(A_k;\eta_k))$ such that

$$\nabla_A E(A_k;\eta_k) \to \sum_{i=1}^{r} \theta_i u_i u_i^T \in \partial \lambda_{\max}(A) \quad \text{as} \quad K \ni k \to \infty. \tag{5.36}$$

For all $k \ge 0$, we have

$$\nabla_A E(A_k;\eta_k) = \sum_{i=1}^{p} \theta_{i,k} u_i(A_k) u_i(A_k)^T \quad \text{and} \quad \theta_{i,k} = \frac{\exp \frac{\lambda_i(A_k) - \lambda_{\max}(A_k)}{\eta_k}}{\sum_{i=1}^{p} \exp \frac{\lambda_i(A_k) - \lambda_{\max}(A_k)}{\eta_k}},$$

where $A_k u_i(A_k) = \lambda_{\max}(A_k) u_i(A_k)$, $\|u_i(A_k)\|_2 = 1$, and the vectors $u_i(A_k)$ are pairwise orthogonal for $i = 1, \ldots, p$; cf. [48, sect. 4]. We have $\sum_{i=1}^{p} \theta_{i,k} = 1$ and $\theta_{i,k} \in [0,1]$. Hence, we can assume w.l.o.g. that for all $i \in \{1, \ldots, p\}$, it holds that $u_i(A_k) \to u_i \in \mathbb{R}^p$, $\theta_{i,k} \to \theta_i \in [0,1]$ as $k \to \infty$, $\|u_i\|_2 = 1$, and $\sum_{i=1}^{p} \theta_i = 1$. We have $A_k u_i(A_k) = \lambda_i(A_k) u_i(A_k)$ for all $k \ge 0$, $A_k \to A$ as $k \to \infty$, and λ is continuous (cf. [33, Cor. 6.3.8]), showing that u_i is an eigenvector of A corresponding to $\lambda_i(A)$. Moreover, $0 = u_i(A_k)^T u_j(A_k) \to u_i^T u_j$ as $k \to \infty$ for all $i \ne j$ implies that u_i are pairwise orthogonal.

Now, let $i \in \{1, \ldots, p\}$ be an index such that $\lambda_i(A) < \lambda_{\max}(A)$, i.e., $i > r(A)$. We obtain that $\lambda_i(A_k) - \lambda_{\max}(A_k) \leq (\lambda_i(A) - \lambda_{\max}(A))/2 < 0$ for all $k \geq 0$ sufficiently large. Hence, $\theta_{i,k} \to 0$ as $k \to \infty$, resulting in $\theta_i = 0$. Moreover, it holds that

$$\operatorname{conv} \left\{ uu^T : \, Au = \lambda_{\max}(A)u, \; \|u\|_2 = 1, \; u \in \mathbb{R}^p \right\} = \partial \lambda_{\max}(A)$$

(cf. [48, sect. 4]), and hence we conclude that (5.36) holds. We have $DA(x^k) \to DA(x)$ as $k \to \infty$, and therefore (5.34) and (5.35) imply the assertion. ∎

We use the notation $(\nu_k, \eta_k)_{\mathbb{N}_0}$ to indicate a sequence distinguishing it from its elements (ν_k, η_k) and to avoid using $((\nu_k, \eta_k))$, and $(\nu_k, \eta_k)_K$ to denote a subsequence of $(\nu_k, \eta_k)_{\mathbb{N}_0}$. In addition to Lemma 5.11, we apply the next result to prove Theorem 5.9.

LEMMA 5.12. *Let the conditions of Theorem* 5.9 *be fulfilled. Moreover, let $\bar{x} \in \mathbb{R}^n$ be given, let $(x^k) \subset \mathbb{R}^n$, and let (ν_k), $(\eta_k) \subset \mathbb{R}_{++}$ be sequences converging to $\bar{x}$ and 0 as $k \to \infty$, respectively. We denote $(\tilde{s}^k, \tilde{\lambda}_k) = (\tilde{s}(x^k; \nu_k, \eta_k), \tilde{\lambda}(x^k; \nu_k, \eta_k))$, where $(\tilde{s}(x; \nu, \eta), \tilde{\lambda}(x; \nu, \eta))$ is defined in* (5.28)*. Then the following conditions hold true:*

1. *The sequence $(\tilde{s}^k, \tilde{\lambda}_k)_{\mathbb{N}_0}$ has a convergent subsequence $(\tilde{s}^k, \tilde{\lambda}_k)_K$. In particular, there exist $(\bar{s}, \bar{\lambda}) \in \mathbb{R}^p \times \mathbb{R}_+$ and $\bar{\alpha}$, $\bar{\beta} \in \mathbb{R}$ such that*

$$\tilde{s}^k = (s(\tilde{\lambda}_k; x^k), \tilde{s}^k_{p+1}, \tilde{s}^k_{p+2}) \to (\bar{s}, \bar{\beta}, \bar{\alpha}) \quad \textit{and} \quad \tilde{\lambda}_k \to \bar{\lambda} \quad \textit{as } K \ni k \to \infty. \tag{5.37}$$

2. *If $\bar{\lambda} > -\lambda_{\min}(H(\bar{x}))$ holds, the easy case occurs for* (5.1) *with $x = \bar{x}$, $(\bar{s}, \bar{\lambda})$ is an optimal primal-dual solution of* (5.1) *for $x = \bar{x}$, and $\bar{\alpha} = 0$.*
3. *If $\bar{\lambda} = -\lambda_{\min}(H(\bar{x}))$ holds, the hard case occurs for* (5.1) *with $x = \bar{x}$. Moreover, let $w_i \in \mathbb{R}^p$, $\|w_i\|_2 = 1$, $i = 1, \ldots, r$, be pairwise orthogonal eigenvectors of $H(\bar{x})$ corresponding to $\lambda_{\min}(H(\bar{x}))$, where $r \in \mathbb{N}$. Then the vectors $(\bar{s} + \gamma_i^+ w_i, \bar{\lambda})$ and $(\bar{s} + \gamma_i^- w_i, \bar{\lambda})$ are optimal primal-dual solutions of* (5.1) *for $x = \bar{x}$, where*

$$\gamma_i^+ = -w_i^T \bar{s} + \sqrt{(w_i^T \bar{s})^2 + \bar{\alpha}^2} \quad \textit{and} \quad \gamma_i^- = -w_i^T \bar{s} - \sqrt{(w_i^T \bar{s})^2 + \bar{\alpha}^2}. \tag{5.38}$$

Proof. 1. Let $k \geq 0$ be arbitrary. We show that $(\tilde{s}^k, \tilde{\lambda}_k)_{\mathbb{N}_0}$ is bounded. Since $\|\tilde{s}^k\|_2 \leq \Delta$ holds, $(\tilde{s}^k)$ is bounded. Lemma 5.8 shows that $\tilde{\lambda}_k = \tilde{\lambda}(x^k; \nu_k, \eta_k) > (E(-H(x^k); \eta))_+$, and hence (5.12) implies

$$\tilde{\lambda}_k > -(\lambda_{\min}(H(x^k)))_+. \tag{5.39}$$

Now, (5.26), Lemma 5.8, and (5.39) yield that

$$\tilde{v}(x^k; \nu_k, \eta_k) = -\tfrac{1}{2} \tilde{g}_{\nu_k}^T(x^k)(\widetilde{H}_{\eta_k}(x^k) + \tilde{\lambda}_k I)^{-1} \tilde{g}_{\nu_k}(x^k) - \tfrac{1}{2}\Delta^2 \tilde{\lambda}_k \leq -\tfrac{1}{2}\Delta^2 \tilde{\lambda}_k \leq 0.$$

The left-hand side of the above inequality converges to $v(\bar{x})$ as $k \to \infty$ by Theorem 5.9 and $\Delta > 0$ holds, implying that $(\tilde{\lambda}_k)$ is bounded. In particular, $(\tilde{s}^k, \tilde{\lambda}_k)_{\mathbb{N}_0}$ is bounded and has a convergent subsequence $(\tilde{s}^k, \tilde{\lambda}_k)_K$. Hence, (5.29) implies that (5.37) holds for some $(\bar{s}, \bar{\lambda}) \in \mathbb{R}^p \times \mathbb{R}_+$ and $\bar{\alpha}$, $\bar{\beta} \in \mathbb{R}$.

Next, (5.14) shows that a necessary optimality condition of (5.22) is

$$\Delta^2 = \|s(\tilde{\lambda}_k; x^k)\|_2^2 + \frac{2\nu_k}{\tilde{\lambda}_k^2} + \frac{2\nu_k}{(\tilde{\lambda}_k - E(-H(x^k); \eta_k))^2} = \|\tilde{s}^k\|_2^2,$$

where we have used (5.30) and (5.30) to establish the second equality. Hence, by applying (5.37) we obtain that

$$\Delta^2 = \|\tilde{s}^k\|_2^2 \to \|\bar{s}\|_2^2 + \bar{\beta}^2 + \bar{\alpha}^2 \quad \text{as} \quad K \ni k \to \infty. \tag{5.40}$$

Moreover, from (5.39), we deduce that

$$H(\bar{x}) + \bar{\lambda} I \succcurlyeq 0 \quad \text{and} \quad \bar{\lambda} \geq 0. \tag{5.41}$$

Using (5.27) and (5.39), we have

$$0 = (H(x^k) + \tilde{\lambda}_k I)s(\tilde{\lambda}_k; x^k) + g(x^k) \to (H(\bar{x}) + \bar{\lambda} I)\bar{s} + g(\bar{x}) \text{ as } K \ni k \to \infty. \tag{5.42}$$

2. Now, we verify that $(\bar{s}, \bar{\lambda})$ is an optimal primal-dual solution of (5.1) for $x = \bar{x}$ and $\bar{\alpha} = 0$ if $\bar{\lambda} > -\lambda_{\min}(H(\bar{x}))$. By assumption $H(\bar{x}) + \bar{\lambda} I$ is invertible, and hence (5.42) implies that $\bar{s}$ is the unique solution to $(H(\bar{x}) + \bar{\lambda} I)\bar{s} = -g(\bar{x})$. Therefore, (5.27) and (5.42) result in $s(\bar{\lambda}; \bar{x}) = \bar{s}$. Moreover, (5.40) implies that $\|\bar{s}\|_2 \leq \Delta$.

If $\bar{\lambda} > 0$, then continuity of $\lambda_{\min}$, $\bar{\lambda} > -\lambda_{\min}(H(\bar{x}))$, $\tilde{\lambda}_k \to \bar{\lambda}$ as $K \ni k \to \infty$, and (5.11) imply that $\tilde{\lambda}_k \geq \bar{\lambda}/2 > 0$ and $\tilde{\lambda}_k - E(-H(x^k); \eta_k) \geq (\bar{\lambda} + \lambda_{\min}(H(\bar{x})))/2 > 0$ for all $k \in K$ sufficiently large. Therefore, we obtain from (5.30) that

$$\tilde{s}^k_{p+1} = \frac{\sqrt{2\nu_k}}{\tilde{\lambda}_k} \to 0 \text{ and } \tilde{s}^k_{p+2} = \frac{\sqrt{2\nu_k}}{\tilde{\lambda}_k - E(-H(x^k); \eta_k)} \to 0 \text{ as } K \ni k \to \infty, \tag{5.43}$$

and therefore $\bar{\alpha}$, $\bar{\beta} = 0$. Now, (5.40) implies that $\Delta^2 = \|\bar{s}\|_2^2$.

Hence, $(s(\bar{\lambda}; \bar{x}), \bar{\lambda})$ satisfies $\bar{\lambda}(\|\bar{s}\|_2^2 - \Delta^2) = 0$, and therefore it fulfills (5.3), implying that it is an optimal primal-dual solution of (5.1) for $x = \bar{x}$ by Theorem 5.1. Theorem 5.1 further implies that the easy case occurs.

3. Next, we establish that the vectors $(\bar{s} + \gamma_i^+ w_i, \bar{\lambda})$ and $(\bar{s} + \gamma_i^- w_i, \bar{\lambda})$ are optimal primal-dual solutions of (5.1) for $x = \bar{x}$ if $\bar{\lambda} = -\lambda_{\min}(H(\bar{x}))$. Let $i \in \{1, \ldots, r\}$ be arbitrary. The numbers γ_i^+ and γ_i^- solve

$$\gamma_i^2 + 2\gamma_i w_i^T \bar{s} - \bar{\alpha}^2 = 0.$$

Using $\|w_i\|_2 = 1$ and (5.40), we obtain for $\gamma_i \in \{\gamma_i^-, \gamma_i^+\}$ that

$$\|\bar{s} + \gamma_i w_i\|_2^2 = \|\bar{s}\|_2^2 + 2\gamma_i w_i^T \bar{s} + \gamma_i^2 = \Delta^2 - \bar{\alpha}^2 - \bar{\beta}^2 + 2\gamma_i w_i^T \bar{s} + \gamma_i^2 \leq \Delta^2 \tag{5.44}$$

with equality if $\bar{\beta} = 0$ and, moreover, (5.42) and $(H(\bar{x}) + \bar{\lambda} I)w_i = 0$ results in

$$(H(\bar{x}) + \bar{\lambda} I)(\bar{s} + \gamma_i w_i) = (H(\bar{x}) + \bar{\lambda} I)\bar{s} = -g(\bar{x}). \tag{5.45}$$

If $\bar{\lambda} > 0$, (5.43) shows that $\bar{\beta} = 0$. Hence, (5.44) implies that $\bar{\lambda}(\|\bar{s} + \gamma_i w_i\|_2^2 - \Delta^2) = 0$.

Moreover, (5.41), (5.42), (5.44), and (5.45) and the above complementarity condition yield that $(\bar{s} + \gamma_i w_i, \bar{\lambda})$, $\gamma_i \in \{\gamma_i^-, \gamma_i^+\}$, fulfill (5.3) and, hence, are optimal primal-dual solutions of (5.1) for $x = \bar{x}$ by Theorem 5.1. Theorem 5.1 further implies that the hard case occurs. ∎

The proof of Theorem 5.9 requires the gradient of $\wp$ (see (5.33)), which is given by

$$\nabla_x \wp(x, s) = \nabla_x g(x)^T s + \tfrac{1}{2}\nabla_x s^T H(x) s = \nabla_x g(x)^T s + \tfrac{1}{2} DH(x)^*[ss^T]. \tag{5.46}$$

Indeed, the first equality in (5.46) follows from the chain rule, and the second by using a similar derivation as in (4.7).

Proof of Theorem 5.10. 1. Let $\bar{x} \in \mathbb{R}^n$ be arbitrary. The function v is locally Lipschitz continuous (cf. [27, Thm. 4.1]), and hence $\partial v(\bar{x})$ is well-defined. From (5.1), (5.33), and [21, Thm. 2.1], we have

$$\partial v(\bar{x}) = \operatorname{conv} \{ \nabla_x \wp(\bar{x}, s^*) : s^* \in \mathcal{S}^*_{\mathrm{TR}}(\bar{x}) \}, \tag{5.47}$$

where $\mathcal{S}^*_{\mathrm{TR}}(\bar{x})$ denotes the set of optimal solutions of (5.1) for $x = \bar{x}$.

Next, we establish that gradient consistency holds, i.e., that (3.5) holds, distinguishing whether the easy or the hard case occurs for (5.1) with $x = \bar{x}$. The inclusion $\partial v(\bar{x}) \subset S_{\tilde{v}}(\bar{x})$ follows from v being locally Lipschitz continuous, where $S_{\tilde{v}}(x)$ is defined in (3.4); cf. Lemma 3.2. Let $z \in \mathbb{R}^p$ be such that there exist sequences $(x^k) \subset \mathbb{R}^n$ and (ν_k), $(\eta_k) \subset \mathbb{R}_{++}$ fulfilling $x^k \to \bar{x}$ and ν_k, $\eta_k \to 0$, and

$$\nabla_x \tilde{v}(x^k; \nu_k, \eta_k) \to z \quad \text{as} \quad k \to \infty. \tag{5.48}$$

Lemma 5.12 implies that the sequence $(\tilde{s}^k, \tilde{\lambda}_k)_{\mathbb{N}_0}$ of optimal primal-dual solutions $(\tilde{s}^k, \tilde{\lambda}_k)$ of (5.22) for $(x, \nu, \eta) = (x^k, \nu_k, \eta_k)$ has a convergent subsequence $(\tilde{s}^k, \tilde{\lambda}_k)_K$. Moreover, the sequence $(s(\tilde{\lambda}_k; x^k), \tilde{\lambda}_k)_K$ converges to $(\bar{s}, \bar{\lambda})$ and $\tilde{s}_{p+2} \to \bar{\alpha}$ as $K \ni k \to \infty$, where $\bar{s} \in \mathbb{R}^p$, $\bar{\lambda} \geq 0$, and $\bar{\alpha} \in \mathbb{R}$, and $s(\lambda; x)$ is defined in (5.27).

In addition, Lemma 5.11 applies with $A = -H$ and shows that there exists a subsequence $(\nabla_x(E(-H(x^k); \eta_k)))_{K'}$ of $(\nabla_x(E(-H(x^k); \eta_k)))_K$ such that

$$\nabla_x(E(-H(x^k); \eta_k)) \to -\sum_{i=1}^{r} \theta_i DH(\bar{x})^*[w_i w_i^T] \quad \text{as} \quad K' \ni k \to \infty, \tag{5.49}$$

where $1 \leq r \leq r(A(\bar{x}))$, $r(A(\bar{x}))$ denotes the multiplicity of $\lambda_{\max}(A(\bar{x}))$, $\theta_i \in [0, 1]$, and $\sum_{i=1}^{r} \theta_i = 1$. Moreover, $\|w_i\|_2 = 1$ and w_i are pairwise orthogonal eigenvectors of $A(\bar{x}) = -H(\bar{x})$ corresponding to $\lambda_{\max}(A(\bar{x})) = -\lambda_{\min}(H(\bar{x}))$. We define $r = r(A(\bar{x}))$.

Hence, (5.32), (5.49), and g and H being continuously differentiable show that

$$\nabla_x \tilde{v}(x^k; \nu_k, \eta_k) \to \nabla_x \wp(\bar{x}, \bar{s}) + (\bar{\alpha}^2/2) \sum_{i=1}^{r} \theta_i DH(\bar{x})^*[w_i w_i^T] \text{ as } K' \ni k \to \infty. \tag{5.50}$$

If the easy case occurs for (5.1) with $x = \bar{x}$, Lemma 5.12 implies that $\bar{s} \in \mathcal{S}^*_{\mathrm{TR}}(\bar{x})$ and $\bar{\alpha} = 0$. By applying (5.47), (5.48), and (5.50), we deduce that $z \in \partial v(\bar{x})$.

If the hard case occurs for (5.1), Lemma 5.12 further implies that $\bar{s} + \gamma_i^+ w_i$ and $\bar{s} + \gamma_i^- w_i$ are optimal solutions of (5.1) for $x = \bar{x}$, where γ_i^+ and γ_i^- are defined in (5.38). If $\bar{\alpha} = 0$, (5.38) implies that either γ_i^+ or γ_i^- is zero, and hence $\bar{s}$ is an optimal solution of (5.1) for $x = \bar{x}$, and hence (5.47), (5.48), and (5.50) imply that $z \in \partial v(\bar{x})$. If $\bar{\alpha} > 0$, (5.38) results in $\gamma_i^+ - \gamma_i^- = 2\sqrt{(w_i^T \bar{s})^2 + \bar{\alpha}^2} > 0$. We define

$$\tau_i^+ = \frac{-\gamma_i^-}{\gamma_i^+ - \gamma_i^-} \quad \text{and} \quad \tau_i^- = \frac{\gamma_i^+}{\gamma_i^+ - \gamma_i^-}. \tag{5.51}$$

Furthermore, (5.38) implies that $\gamma_i^+ > 0$ and $\gamma_i^- < 0$, and hence (5.51) shows that

$$\begin{aligned} &\tau_i^+ > 0, \quad \tau_i^- > 0, \quad \tau_i^+ + \tau_i^- = 1, \\ &\tau_i^+ \gamma_i^+ + \tau_i^- \gamma_i^- = \frac{-\gamma_i^- \gamma_i^+ + \gamma_i^+ \gamma_i^-}{\gamma_i^+ - \gamma_i^-} = 0, \quad \text{and} \quad \tau_i^+ (\gamma_i^+)^2 + \tau_i^- (\gamma_i^-)^2 = \bar{\alpha}^2. \end{aligned} \tag{5.52}$$

Using (5.33) and (5.46), we obtain for $\gamma_i \in \{\gamma_i^-, \gamma_i^+\}$ that

$$\begin{aligned} \nabla_x \wp(\bar{x}, \bar{s} + \gamma_i w_i) = \nabla_x g(\bar{x})^T \bar{s} &+ (1/2) DH(\bar{x})^*[\bar{s}\bar{s}^T] + \gamma_i \nabla_x g(\bar{x})^T w_i \\ &+ (1/2)\gamma_i DH(\bar{x})^*[w_i \bar{s}^T + \bar{s} w_i^T] + (1/2)(\gamma_i)^2 DH(\bar{x})^*[w_i w_i^T], \end{aligned}$$

resulting in

$$\begin{aligned}&\tau_i^+\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^+w_i)+\tau_i^-\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^-w_i)\\&\quad=(\tau_i^-+\tau_i^+)\nabla_x g(\bar{x})^T\bar{s}+(1/2)(\tau_i^-+\tau_i^+)DH(\bar{x})^*[\bar{s}\bar{s}^T]\\&\qquad+(\tau_i^+\gamma_i^++\tau_i^-\gamma_i^-)\nabla g(\bar{x})^Tw_i+(1/2)(\tau_i^+\gamma_i^++\tau_i^-\gamma_i^-)DH(\bar{x})^*[w_i\bar{s}^T+\bar{s}w_i^T]\\&\qquad+(1/2)(\tau_i^+(\gamma_i^+)^2+\tau_i^-(\gamma_i^-)^2)DH(\bar{x})^*[w_iw_i^T].\end{aligned}$$

Hence, (5.52) implies that

$$\begin{aligned}&\tau_i^+\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^+w_i)+\tau_i^-\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^-w_i)\\&\quad=\nabla_x g(\bar{x})^T\bar{s}+(1/2)DH(\bar{x})^*[\bar{s}\bar{s}^T]+(\bar{\alpha}^2/2)DH(\bar{x})^*[w_iw_i^T],\end{aligned}$$

implying, with $\sum_{i=1}^r\theta_i=1$ and (5.46), that

$$\begin{aligned}&\sum_{i=1}^r\theta_i\tau_i^+\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^+w_i)+\sum_{i=1}^r\theta_i\tau_i^-\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^-w_i)\\&\quad=\nabla_x\wp(\bar{x},\bar{s})+(\bar{\alpha}^2/2)\sum_{i=1}^r\theta_i DH(\bar{x})^*[w_iw_i^T].\end{aligned}\tag{5.53}$$

Moreover, using (5.52), we have $\sum_{i=1}^r\theta_i\tau_i^++\sum_{i=1}^r\theta_i\tau_i^-=\sum_{i=1}^r\theta_i(\tau_i^++\tau_i^-)=1$. The limit in (5.50) equals (5.53). Now, we use the fact that $\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^+w_i)$ and $\nabla_x\wp(\bar{x},\bar{s}+\gamma_i^-w_i)$ are contained in $\partial v(\bar{x})$ (cf. Lemma 5.12), implying that (5.53) is a convex combination of elements of $\partial v(\bar{x})$. Hence, (5.47), (5.48), and (5.50) yield $z\in\partial v(\bar{x})$.

2. Adapting the above reasoning and using (5.32), we obtain that $(\nabla_x\tilde{v}(x^k;\nu_k,\eta_k))$ has a converging subsequence if $x^k\to x$ and ν_k, $\eta_k\to 0^+$ as $k\to\infty$. ∎

Theorems 5.9 and 5.10 imply that the function $\tilde{\psi}_j:\mathbb{R}^n\times\mathbb{R}^2_{>0}\to\mathbb{R}$ given by

$$\tilde{\psi}_j(x;\nu,\eta)=h_j(x)-\min_{\tilde{s}\in\mathbb{R}^{p+2}}\left\{(1/2)\tilde{s}^T\widetilde{H}_{\eta,j}(x)\tilde{s}+\widetilde{g}_{\nu,j}(x)^T\tilde{s}:\ \ \|\tilde{s}\|_2\le\Delta\right\}\tag{5.54}$$

is a smoothing function of ψ_j (see (1.6)), where $h_j(x)=a_j(x)$, $g_j(x)=-\bar{\Sigma}^{1/2}b_j(x)$, and $H_j(x)=-\bar{\Sigma}^{1/2}C_j(x)\bar{\Sigma}^{1/2}$. Moreover, $\widetilde{H}_{\eta,j}$ and $\widetilde{g}_{\nu,j}$ are defined as in (5.23) with H and g replaced by H_j and g_j, respectively. The representation of $\tilde{\psi}_j$ results from (1.6) being transformed into the TRP (5.1) using $d\mapsto s=\bar{\Sigma}^{-1/2}d$.

THEOREM 5.13. *Let $\bar{\Sigma}\in\mathbb{S}^p_{++}$, and let $a_j:\mathbb{R}^n\to\mathbb{R}$, $b_j:\mathbb{R}^n\to\mathbb{R}^p$, and $C_j:\mathbb{R}^n\to\mathbb{R}^p$ be q-times continuously differentiable, where $q\ge 1$ and $j\in J$. Then the following conditions hold true:*

1. *The function $\tilde{\psi}_j$ defined in (5.54) is a smoothing function of ψ_j, $\tilde{\psi}_j(\cdot;\nu,\eta)$ is q-times continuously differentiable for every ν, $\eta>0$, and gradient consistency holds.*
2. *Let $x\in\mathbb{R}^n$ be given, and let $(x^k)\subset\mathbb{R}^n$ and (ν_k), $(\eta_k)\subset\mathbb{R}_{++}$ be sequences converging to x and 0 as $k\to\infty$, respectively. Then there exists a convergent subsequence $(\nabla_x\tilde{\psi}_j(x^k;\nu_k,\eta_k))_K$ of $(\nabla_x\tilde{\psi}(x^k;\nu_k,\eta_k))$.*

The computational cost of evaluating (5.54) is essentially the same as the evaluation of (1.6) since $\widetilde{H}_{\eta,j}(x)$ (see (5.23)) is a block-diagonal matrix for $x\in\mathbb{R}^n$, implying that our smoothing approach is tractable both theoretically and practically.

6. Convergence of the homotopy method. We show that a sequence of KKT-tuples of (3.2) generated by Algorithm 3.1 converges to a stationary point of the DROP (3.1) under mild assumptions. We define $\widetilde{F}_j : \mathbb{R}^n \times \mathbb{R}^3_{++} \to \mathbb{R}$ by

$$\widetilde{F}_j(x;t) = \tilde{\varphi}_j(x;\tau) + \tilde{\psi}_j(x;\nu,\eta) \tag{6.1}$$

and recall that $F_j : \mathbb{R}^n \to \mathbb{R}$, $F_j(x) = \varphi_j(x) + \psi_j(x)$ for all $j \in J$, where we set $t = (\tau,\nu,\eta)$, and $\tilde{\varphi}_j$ and $\tilde{\psi}_j$ are defined in (4.3) and (5.54), respectively. Suitable assumptions on (1.4) imply that the DROP (3.2) has feasible points.

PROPOSITION 6.1. *Let $z \in \mathbb{R}^n$ be a strictly feasible point for* (3.1)*, and let the conditions of Theorems* 4.2 *and* 5.13 *be fulfilled for any $j \in J \setminus \{0\}$. Then z is a strictly feasible point to* (3.2) *for all sufficiently small $t > 0$.*

Proof. Theorems 4.2 and 5.13 and (6.1) imply that

$$\widetilde{F}_j(z;t) = \tilde{\varphi}_j(z;\tau) + \tilde{\psi}_j(z;\nu,\eta) \to F_j(z) \quad \text{as} \quad t = (\tau,\nu,\eta) \to 0^+$$

for all $j \in J \setminus \{0\}$, establishing the assertion. ∎

Next, we provide a global convergence result of Algorithm 3.1.

THEOREM 6.2. *Let the conditions of Theorems* 4.2 *and* 5.13 *hold for all $j \in J$. Choose $\varepsilon_{\min}$, $t_{\min} = 0$, and let the sequence $(x^k, \vartheta^k)_{\mathbb{N}_0}$ be generated by Algorithm* 3.1*. Then every accumulation point of $(x^k, \vartheta^k)_{\mathbb{N}_0}$ is a KKT-point of* (3.1).

Proof. Let $(\bar{x}, \bar{\vartheta})$ be an accumulation point of $(x^k, \vartheta^k)_{\mathbb{N}_0}$. Then there exists a subsequence $(x^k, \vartheta^k)_K$ of $(x^k, \vartheta^k)_{\mathbb{N}_0}$ converging to $(\bar{x}, \bar{\vartheta})$ as $K \ni k \to \infty$. Further, it holds that $0 \le \chi(x^k, \vartheta^k; t^k) \le \varepsilon_k$ for all $k \ge 0$, where χ is defined in (3.3). Since $\varepsilon_k \to 0$ as $k \to \infty$, we obtain from (6.1) and Theorems 4.2 and 5.13 that

$$\varepsilon_k \ge |\min\{-\widetilde{F}_j(x^k;t^k), \vartheta^k_j\}| \to |\min\{-F_j(\bar{x}), \bar{\vartheta}_j\}| = 0 \quad \text{as} \quad K \ni k \to \infty \quad \text{for all } j \in J\setminus\{0\}.$$

Because $(a,b) \mapsto \min\{a,b\}$ is a complementarity function, we have $\bar{\vartheta}_j F_j(\bar{x}) = 0$, $F_j(\bar{x}) \le 0$, and $\bar{\vartheta}_j \ge 0$ for all $j \in J \setminus \{0\}$. We can assume w.l.o.g. that the sequences $(\nabla_x \tilde{\varphi}_j(x^k;\tau_k))_K$, $j \in J$, and $(\nabla_x \tilde{\psi}_j(x^k;\nu_k,\eta_k))_K$, $j \in J$, are convergent; cf. Theorems 4.2 and 5.13. Hence, there exist v_j, $w_j \in \mathbb{R}^n$ such that

$$\nabla_x \tilde{\varphi}_j(x^k;\tau_k) \to v_j, \quad \nabla_x \tilde{\psi}_j(x^k;\nu_k,\eta_k) \to w_j \quad \text{as} \quad K \ni k \to \infty \quad \text{for all} \quad j \in J.$$

Now, let $j \in J$ be arbitrary. We verify that $v_j + w_j \in \partial F_j(\bar{x})$. Theorems 4.2 and 5.13 apply and yield that $v_j \in \partial \varphi_j(\bar{x})$ and $w_j \in \partial \psi_j(\bar{x})$ due to gradient consistency. Next, [21, Thm. 2.1] and [22, Prop. 2.3.6] show that φ_j and ψ_j are regular according to [22, Def. 2.3.4], and therefore [22, Cor. 3, p. 40] results in $\partial F_j(\bar{x}) = \partial \varphi_j(\bar{x}) + \partial \psi_j(\bar{x})$, showing that $v_j + w_j \in \partial F_j(\bar{x})$. Hence, we have

$$v_0 + w_0 + \sum_{j \in J\setminus\{0\}} \bar{\vartheta}_j (v_j + w_j) \in \partial F_0(\bar{x}) + \sum_{j \in J\setminus\{0\}} \bar{\vartheta}_j \partial F_j(\bar{x}).$$

Moreover, $\chi(x^k, \vartheta^k; t^k) \to 0$ as $k \to \infty$, where χ is defined in (3.3), implies that

$$\nabla_x \widetilde{F}_0(x^k;t^k) + \sum_{j \in J\setminus\{0\}} (\vartheta^k)_j \nabla_x \widetilde{F}_j(x^k;t^k) \to 0 \quad \text{as} \quad K \ni k \to \infty,$$

and therefore we deduce that $0 \in \partial F_0(\bar{x}) + \sum_{j \in J\setminus\{0\}} \bar{\vartheta}_j \partial F_j(\bar{x})$. ∎

If we only assume (x^k) to have a convergent subsequence, we need to impose a suitable CQ for (3.1) to deduce convergence of a subsequence of (ϑ^k); cf. [64, Thm. 3.2]. Moreover, the existence of KKT-tuples of the DROP (3.2) may be verified under suitable CQs for (3.1); cf. [64].

7. Numerical examples. We construct DROPs from the test set of Moré, Garbow, and Hillstrom [46] consisting of standard NLPs modeling design variables as uncertain, which for the case of RO has been considered in, e.g., [8, 39]:

$$\min_{x\in\mathbb{R}^n} \sup_{P\in\mathcal{P}_\epsilon} \mathbb{E}_P[f_0(x+\xi)], \tag{7.1}$$

where $\mathcal{P}_\epsilon$ is defined as in (1.2) with $\bar{\mu}=0$, $\Delta=\sqrt{\epsilon}$, $\bar{\Sigma}_0=0$, and $\bar{\Sigma}=\bar{\Sigma}_1=\epsilon I$. We consider $\epsilon\in\{10^{-3},10^{-2}\}$ and refer the reader to Appendix A for a description of how we selected test problems. The problem under consideration is

$$\min_{x\in\mathbb{R}^n} F_0(x), \tag{7.2}$$

where F_0 is defined in (6.1) and $a_0(x)=f_0(x)$, $b_0(x)=\nabla f_0(x)$, $C_0(x)=\nabla^2 f_0(x)$ are chosen in (1.5) and (1.6). One goal of our experiments is to show that Algorithm 3.1 is an efficient method to compute stationary points of (7.2). We compare the performance of Algorithm 3.1 with the bundle method `MPBNGC` [43, 44] applied to (7.2) and the nonlinear SDP solver `PENLAB` [28] applied to

$$\begin{aligned} &\min_{x\in\mathbb{R}^n,\,\gamma\in\mathbb{R},\,\lambda\in\mathbb{R}_+,\,y\in\mathbb{R}^p,\,\Lambda,\,\Upsilon\in\mathbb{S}^p_+} && a_0(x)-(1/2)\gamma+(1/2)I\bullet\Upsilon \\ &\quad\text{s.t.}\ \begin{bmatrix}\lambda I+\Upsilon-\Lambda & y\\ y^T & -\lambda\Delta^2-\gamma\end{bmatrix}\succcurlyeq 0, && \begin{array}{c}\operatorname{svec}(\Upsilon-\Lambda+\bar{\Sigma}^{1/2}C_0(x)\bar{\Sigma}^{1/2})=0,\\ y+\bar{\Sigma}^{1/2}b_0(x)=0,\end{array}\end{aligned} \tag{7.3}$$

where $\operatorname{svec}:\mathbb{S}^p\to\mathbb{R}^{p(p+1)/2}$ transforms the lower triangular part of a symmetric matrix into a vector. From (1.2) and (1.8), we deduce that (7.3) is equivalent to (7.2).

A further goal of our numerical tests is to show that stationary points of (7.2) are more robust than those of the nominal problem

$$\min_{x\in\mathbb{R}^n} f_0(x), \tag{7.4}$$

and of a sample average approximation (SAA) of the stochastic program

$$\min_{x\in\mathbb{R}^n} \mathbb{E}_{\bar{P}_\epsilon}[f_0(x+\xi)], \tag{7.5}$$

with $\bar{P}_\epsilon=N(0,(\epsilon/10)I)$, even though we approximate (7.1) by (7.2). We chose $\bar{P}_\epsilon=N(0,(\epsilon/10)I)$ to mimic the setup considered in [23, sect. 4.3].

7.1. Implementation details. We provide implementation details of Algorithm 3.1 and of the application of `MPBNGC` to (7.2) and of `PENLAB` to (7.3). We implemented Algorithm 3.1 in `Julia` [10] using `Ipopt` [61] and its `Julia` interface `Ipopt.jl`. We chose the same stopping criterion for each iteration of Algorithm 3.1. We used the default settings of `Ipopt`, with the exception of modifying the overall termination tolerance `tol`. We computed the gradient of the smoothing functions $\tilde{\varphi}_0$ (see (4.3)) and $\tilde{\psi}_0$ (see (5.54)) based on the formulas (4.6) and (5.32), respectively, and used `Ipopt` with L-BFGS. We chose $\nu_{\min}=10^{-8}$, $\eta_{\min}$, $\tau_{\min}=\sqrt{\nu_{\min}}$, η_0, $\tau_0=\sqrt{\nu_0}$, $\nu_{k+1}=\min\{\rho^2\nu_k,\nu_{\min}\}$, and η_{k+1}, $\tau_{k+1}=\min\{\rho\eta_k,\nu_{\min}\}$, where $\nu_0>0$, $\rho=0.1$.

For $\mathtt{tol} = 10^{-4}$ and $\nu_0 = 0.1$, the above choices of the smoothing parameters are motivated by Theorems 4.2 and 5.6. Evaluating the smoothing function $\widetilde{F}_0$ (see (6.1)) of F_0 (see (7.2)) at (x, t) requires $f_0(x)$ (see (7.4)), $\nabla f_0(x)$, and $\nabla^2 f_0(x)$. To obtain $\nabla_x \widetilde{F}_0(x; t)$, we computed the gradients of $x \mapsto \widetilde{s}^T \nabla^2 f_0(x) \widetilde{s}$, where $\widetilde{s}$ are the first p components of the optimal solution of the TRP (5.54), and of two mapping of the form $x \mapsto \nabla^2 f_0(x) \bullet R$, where $R \in \mathbb{S}^p$; cf. (4.6) and (5.32). To initialize the solution of the smoothed problem (3.2) in the $(k+1)$st iteration of Algorithm 3.1, we used the approximate stationary point obtained in the kst iteration.

For the application of `MPBNGC`, we used the same setup as in [43, sect. 6], except that we set different termination tolerances and $\mathtt{GAM} = 0.5$ for all test problems. We implemented a `Julia` interface for `MPBNGC`.[1] We exploited the regularity of φ_0 (see (1.5)) and of ψ_0 (see (1.6)) and computed subgradients of the objective F_0 of (7.2) via the sum of a subgradient of φ_0 and of ψ_0, which we evaluated based on (4.9) and (5.47).

For `PENLAB`, we computed derivatives of f_0 up to fourth order using the automatic differentiation tool `ADiGator` [62]. We excluded the test function mgh35 from the tests with `PENLAB` as `ADiGator` does not support automatic differentiation for it. We chose the same initial values for x that we passed to Algorithm 3.1. We obtained the remaining initial points for the variables in (7.3) by applying `PENLAB` to (7.3) for fixed x. We used `PENLAB`'s default settings expect for the stopping criteria `outer_stop_limit` and `kkt_stop_limit`, which we chose as equal.

We scaled f_0 using the gradient scaling of `Ipopt` (see [61, sect. 3.8]) and chose x_N^* as initial value for Algorithm 3.1 and `MPBNGC`, where x_N^* is the stationary point computed by `Ipopt` for the nominal problem (7.4) with termination tolerance 10^{-5} and the above settings. The TRPs (1.6) and (5.54) were solved using [47, Alg. 3.14]. For the `Julia` codes, derivatives of f_0 were obtained with the automatic differentiation package `ForwardDiff` [52]. We took advantage of the fact that the DROP (7.1) models uncertain decision variables and used the fact that $\nabla f_0 = \nabla a_0 = b_0$ and $\nabla b_0 = C_0$.

7.2. Comparison of homotopy method with `MPBNGC` and `PENLAB`. We compare the performance of Algorithm 3.1 with `MPBNGC` and `PENLAB` in terms of evaluations of f_0 and its derivatives. The stopping criteria of `Ipopt` (see [61, sect. 2.1]), `MPBNGC` (see [44, sect. 3.3]), and `PENLAB` (see [28, Alg. 1]) are different. To be able to make a fair comparison, we applied `Ipopt` to each nominal problem (7.4) with known exact solution (see [46, sect. 3]) and computed the median of the absolute errors of the final objective function values returned by `Ipopt` and the true ones with $\mathtt{tol} = 10^{-2}, 10^{-4}$. Then we applied `MPBNGC` and `PENLAB` to the same problems with termination tolerances $10^{-1}, 10^{-2}, \ldots, 10^{-10}$, and from this list computed the largest ones such that we obtained the same order of magnitude of the errors as with `Ipopt` for the tolerances $10^{-2}, 10^{-4}$. The corresponding criteria for `MPBNGC` are $10^{-4}, 10^{-8}$ and $10^{-2}, 10^{-5}$ for `PENLAB`. This type of "calibration" tries to ensure that stationary points obtained via Algorithm 3.1, `MPBNGC`, and `PENLAB` are of similar accuracy.

We report the median number of evaluations of f_0, of its derivatives, and of the derivatives $\nabla(d^T \nabla^2 f_0(\cdot) d)$, $d \in \mathbb{R}^p$, and $\nabla(\nabla^2 f_0(\cdot) \bullet R)$, $R \in \mathbb{S}^p$, used by Algorithm 3.1, `MPBNGC`, and `PENLAB` with $\epsilon = 10^{-3}$ in Table 1. For each selected test problem, the number of evaluations used in Algorithm 3.1 is the sum of all evaluations of the inner iterations. We chose initial smoothing parameters $\nu_0 = 10^{-1}$, 10^{-3}, and $\eta_0, \tau_0 = \sqrt{\nu_0}$. Moreover, we used the termination tolerances $\mathtt{tol} = 10^{-2}$, 10^{-4} for Algorithm 3.1,

[1] The interface is available from https://github.com/milzj/MPBNGCInterface.jl.

TABLE 1

Median number of evaluations of derivatives of f_0, $x \mapsto d^T\nabla^2 f_0(x)d$, $d \in \mathbb{R}^p$, and of $x \mapsto \nabla^2 f_0(x) \bullet R$, $R \in \mathbb{S}^p$ used by Algorithm 3.1, `MPBNGC`, and `PENLAB` with $\epsilon = 10^{-3}$. Each number is rounded to its nearest integer. `HM(tol`, ν_0`)` refers to Algorithm 3.1 with termination tolerance `tol` and initial smoothing parameter ν_0.

Method	$\#f_0$, $\#\nabla f_0$, $\#\nabla^2 f_0$	$\#\nabla(d^T\nabla^2 f_0 d)$	$\#\nabla(\nabla^2 f_0 \bullet R)$	$\#D^3 f_0$	$\#D^4 f_0$
`HM`(10^{-2}, 10^{-1})	76	15	30	0	0
`HM`(10^{-2}, 10^{-3})	61	15	30	0	0
`MPBNGC`	24	24	24	0	0
`PENLAB`	54	0	0	29	23
`HM`(10^{-4}, 10^{-1})	120	37	75	0	0
`HM`(10^{-4}, 10^{-3})	101	32	64	0	0
`MPBNGC`	69	69	69	0	0
`PENLAB`	88	0	0	59	46

and for `MPBNGC` and `PENLAB` the corresponding ones as stated above. Instead of evaluating the gradient of $x \mapsto d^T\nabla^2 f_0(x)d$, and of $x \mapsto \nabla^2 f_0(x) \bullet R$ for two $R \in \mathbb{S}^p$, we could have computed $D^3 f_0$ once.

Table 1 indicates that Algorithm 3.1 requires about half as many gradient evaluations of $\widetilde{F}_0$ as `MPBNGC` requires subgradient evaluations of F_0. `PENLAB` requires, as opposed to Algorithm 3.1 and `MPBNGC`, third and fourth derivatives of f_0. Hence, `PENLAB` is the most expensive method in terms of evaluations of f_0 and of its derivatives. Table 1 indicates that small initial smoothing parameters can be beneficial, as they result in fewer evaluations.

Our homotopy method requires more f_0-evaluations than `MPBNGC`. We found that this is caused by the line search of `Ipopt` in cases where $\widetilde{F}_0$-changes are so small that they reach the computational accuracy to which we compute $\widetilde{F}_0$ (empirically about the square root of the machine precision). Experiments with a modified line search showed that we can reduce the median number of f_0-, $\nabla(d^T\nabla^2 f_0(\cdot)d)$-, and $\nabla(\nabla^2 f_0(\cdot) \bullet R)$-evaluations of our method by 16%, and then the computational costs of our method are lower than those of `MPBNGC`.

7.3. Details on performance of homotopy method. We discuss the performance of Algorithm 3.1 as a smoothing method with `tol` $= 10^{-4}$ and $\nu_0 = 0.1$. Table 2 lists for mgh01 and mgh03 the number of inner and outer iterations. Moreover, it displays the KKT-error, the distance of the stationary point of the current iteration to that of the previous iteration, and the smoothing parameter ν_k for each outer iteration k of Algorithm 3.1. We deduce that empirically the distance of subsequent stationary points (3.2) converges to zero and that the number of inner iterations decreases monotonically, indicating that the homotopy method is efficient.

The solution of the TRPs (5.54) using [47, Alg. 3.14] for all iterations of Algorithm 3.1 required fewer than six iterations. The evaluation of (1.5) using (4.2) instead of applying SDP solvers is about three orders of magnitudes faster, e.g., for $p = 20$, the quotient of the median run time over 100 randomly generated SDPs of the form (1.5) using (4.2) and `SCS` [50] is $4.340 \cdot 10^{-4}$.

7.4. Comparison of stationary points. For each selected problem, we compare the stationary points x^*_{DR} of (7.2) computed with Algorithm 3.1 using `tol` $=$ 10^{-4} and $\nu_0 = 0.1$, x^*_N of (7.4) and x^*_S of an SAA of (7.5) using the following two

Table 2

For each outer iteration of Algorithm 3.1 applied to (7.1) and $\epsilon = 10^{-3}$, the number of iterations required to compute a stationary point of (3.2), the final KKT-error, relative distance of the initial point and the stationary point, and value of the smoothing parameter ν_k.

Problem	k	#-iter	KKT-error	$\frac{\lVert x^k - x^{k-1}\rVert_2}{\max\{1, \lVert x^{k-1}\rVert_2\}}$	ν_k
mgh01	0	17	$2.272 \cdot 10^{-7}$	0.3329	0.1
	1	10	$2.756 \cdot 10^{-6}$	$9.633 \cdot 10^{-2}$	$1.0 \cdot 10^{-3}$
	2	2	$4.34 \cdot 10^{-5}$	$5.013 \cdot 10^{-5}$	$1.0 \cdot 10^{-5}$
	3	2	$6.355 \cdot 10^{-5}$	$3.975 \cdot 10^{-5}$	$1.0 \cdot 10^{-7}$
	4	2	$9.689 \cdot 10^{-5}$	$3.345 \cdot 10^{-5}$	$1.0 \cdot 10^{-8}$
mgh03	0	25	$7.654 \cdot 10^{-5}$	0.9994	0.1
	1	7	$6.938 \cdot 10^{-7}$	$5.542 \cdot 10^{-5}$	$1.0 \cdot 10^{-3}$
	2	7	$2.303 \cdot 10^{-7}$	$5.495 \cdot 10^{-6}$	$1.0 \cdot 10^{-5}$
	3	5	$4.997 \cdot 10^{-7}$	$5.494 \cdot 10^{-7}$	$1.0 \cdot 10^{-7}$
	4	5	$1.015 \cdot 10^{-6}$	$4.07 \cdot 10^{-8}$	$1.0 \cdot 10^{-8}$

quantities:

$$V_{\mathbb{E}}(x) = \max_{1\le i\le 10} \mathbb{E}_{P_i}[f_0(x+\xi_i)] \quad \text{and} \quad V_{\mathrm{StD}}(x) = \max_{1\le i\le 10} \mathrm{StD}_{P_i}[f_0(x+\xi_i)], \tag{7.6}$$

where $P_i = N(\mu_i, \sigma_i^2 I) \in \mathcal{P}_\epsilon$, and μ_i and σ_i are independent and uniformly distributed on $\{\, \mu \in \mathbb{R}^p : \lVert\mu\rVert_2 \le \Delta \,\}$ and $\{\, \sigma \in \mathbb{R} : 0 \le \sigma^2 \le \epsilon \,\}$, respectively. Here, StD denotes the standard deviation. We approximated expected values using Monte Carlo with 1000 independent samples. The quantities in (7.6) mimic the maximum mean and standard deviations of repeated implementations of x, and $V_{\mathbb{E}}$ is a lower bound on the objective function of (7.1). We computed the stationary points x_N^* and x_S^* using `Ipopt` with `tol` $= 10^{-5}$ and exact Hessian information for nominal and stochastic programs. Tables 3 and 4 display $V_{\mathbb{E}}(x)$ and $V_{\mathrm{StD}}(x)$ for $x \in \{x_{DR}^*, x_N^*, x_S^*\}$ and $\epsilon \in \{10^{-3}, 10^{-2}\}$. In most cases, the distributionally robust stationary point has lower mean and standard deviation than nominal and stochastic stationary points.

Problems mgh33 and mgh34 are quadratic w.r.t. ξ (cf. [46, sect. 3]), and hence the approximation scheme is exact, i.e., (7.1) is equivalent to (7.2). For problems mgh10, mgh11, and mgh17, we obtained very different orders of magnitude of $V_{\mathbb{E}}(x)$ and of $V_{\mathrm{StD}}(x)$ for $x \in \{x_N^*, x_{DR}^*, x_S^*\}$, resulting from exponential terms in the corresponding objective functions; cf. [46, sect. 3].

Table 5 lists the median number of corresponding objective function, gradient, and Hessian evaluations used by `Ipopt` to compute a stationary point of (7.2) using Algorithm 3.1, of (7.4), and of the sample average approximation of (7.5).

8. Conclusion and outlook. We have provided a new algorithmic scheme for both DRO and RO. The main advantages of our approach are that the number of constraints of the DROP is the same as for the nominal problem, MPCCs, and NSDPs are avoided, and any NLP solver can be used to compute stationary points of the DROPs in Algorithm 3.1. Moreover, it is applicable to a large class of problems without the need to implement further algorithms. The numerical experiments indicate that our smoothing method is competitive in comparison with other approaches.

Appendix A. We selected problems from the Moré–Garbow–Hillstrom test set [46], which is available in `Julia` through the package `NLSProblems.jl` (version as of November 16, 2018) using its default setup as follows: We compute for each test

TABLE 3
Quantities $V_{\mathbb{E}}$ and V_{StD} (see (7.6)) evaluated at x_N^*, x_{DR}^*, x_S^* for $\epsilon = 10^{-3}$.

Problem	$V_{\mathbb{E}}(x_N^*)$	$V_{\mathbb{E}}(x_{DR}^*)$	$V_{\mathbb{E}}(x_S^*)$	$V_{\mathrm{StD}}(x_N^*)$	$V_{\mathrm{StD}}(x_{DR}^*)$	$V_{\mathrm{StD}}(x_S^*)$
mgh01	0.1867	0.1536	0.1761	0.2559	0.1528	0.2399
mgh03	$3.175 \cdot 10^6$	$3.135 \cdot 10^1$	$2.899 \cdot 10^1$	$4.507 \cdot 10^6$	$8.083 \cdot 10^1$	$7.328 \cdot 10^1$
mgh04	$3.756 \cdot 10^8$	$3.754 \cdot 10^8$	$3.754 \cdot 10^8$	$5.256 \cdot 10^8$	$5.246 \cdot 10^8$	$5.252 \cdot 10^8$
mgh06	$1.884 \cdot 10^2$	$1.798 \cdot 10^2$	$1.863 \cdot 10^2$	$1.076 \cdot 10^2$	$8.388 \cdot 10^1$	$1.028 \cdot 10^2$
mgh07	0.1778	0.1778	0.1779	0.2089	0.2086	0.209
mgh10	$9.626 \cdot 10^{10}$	$1.356 \cdot 10^6$	$2.134 \cdot 10^6$	$1.303 \cdot 10^{11}$	$3.482 \cdot 10^5$	$1.993 \cdot 10^6$
mgh11	$6.237 \cdot 10^{278}$	$3.283 \cdot 10^1$	$2.258 \cdot 10^{133}$	∞	0.8311	$7.134 \cdot 10^{134}$
mgh13	$4.387 \cdot 10^{-2}$	$4.387 \cdot 10^{-2}$	$4.385 \cdot 10^{-2}$	$5.662 \cdot 10^{-2}$	$5.662 \cdot 10^{-2}$	$5.66 \cdot 10^{-2}$
mgh14	0.7525	0.7492	0.752	0.7223	0.7144	0.7229
mgh17	$7.9421 \cdot 10^{17}$	1.133	$1.735 \cdot 10^{11}$	$2.19 \cdot 10^{19}$	$3.551 \cdot 10^{-2}$	$4.959 \cdot 10^{12}$
mgh20	0.1318	0.1291	0.1309	0.1461	0.1425	0.1453
mgh21	3.92	3.19	3.702	1.723	1.045	1.621
mgh22	0.2164	0.2163	0.2163	0.1219	0.1219	0.1219
mgh25	0.3078	0.3078	0.3073	0.6784	0.6784	0.6768
mgh27	$4.855 \cdot 10^{-2}$	$4.853 \cdot 10^{-2}$	$4.85 \cdot 10^{-2}$	$6.864 \cdot 10^{-2}$	$6.854 \cdot 10^{-2}$	$6.859 \cdot 10^{-2}$
mgh30	0.1408	0.1406	0.1408	$7.52 \cdot 10^{-2}$	$7.485 \cdot 10^{-2}$	$7.519 \cdot 10^{-2}$
mgh31	0.194	0.1924	0.1936	$9.437 \cdot 10^{-2}$	$8.959 \cdot 10^{-2}$	$9.399 \cdot 10^{-2}$
mgh33	$4.514 \cdot 10^2$	$4.514 \cdot 10^2$	$4.508 \cdot 10^2$	$6.369 \cdot 10^2$	$6.369 \cdot 10^2$	$6.365 \cdot 10^2$
mgh34	$2.394 \cdot 10^2$	$2.394 \cdot 10^2$	$2.391 \cdot 10^2$	$3.203 \cdot 10^2$	$3.203 \cdot 10^2$	$3.204 \cdot 10^2$
mgh35	$6.772 \cdot 10^{-2}$	$5.266 \cdot 10^{-2}$	0.1244	0.3531	$2.726 \cdot 10^{-2}$	1.383

TABLE 4
Quantities $V_{\mathbb{E}}$ and V_{StD} (see (7.6)) evaluated at x_N^*, x_{DR}^*, x_S^* for $\epsilon = 10^{-2}$.

Problem	$V_{\mathbb{E}}(x_N^*)$	$V_{\mathbb{E}}(x_{DR}^*)$	$V_{\mathbb{E}}(x_S^*)$	$V_{\mathrm{StD}}(x_N^*)$	$V_{\mathrm{StD}}(x_{DR}^*)$	$V_{\mathrm{StD}}(x_S^*)$
mgh01	1.866	0.7566	1.174	2.581	0.5774	1.548
mgh03	$3.178 \cdot 10^7$	$2.829 \cdot 10^3$	$2.81 \cdot 10^3$	$4.511 \cdot 10^7$	$7.443 \cdot 10^3$	$7.414 \cdot 10^3$
mgh04	$3.752 \cdot 10^9$	$3.728 \cdot 10^9$	$3.744 \cdot 10^9$	$5.246 \cdot 10^9$	$5.142 \cdot 10^9$	$5.233 \cdot 10^9$
mgh06	$1.903 \cdot 10^3$	$8.186 \cdot 10^2$	$1.528 \cdot 10^3$	$5.762 \cdot 10^3$	$2.055 \cdot 10^3$	$4.626 \cdot 10^3$
mgh07	1.793	1.781	1.792	2.129	2.09	2.125
mgh10	$9.616 \cdot 10^{11}$	$9.616 \cdot 10^{11}$	$3.502 \cdot 10^6$	$1.294 \cdot 10^{12}$	$1.294 \cdot 10^{12}$	$3.261 \cdot 10^6$
mgh11	$8.08 \cdot 10^{256}$	$3.283 \cdot 10^1$	$7.044 \cdot 10^{129}$	∞	0.7971	$2.228 \cdot 10^{131}$
mgh13	0.4413	0.4413	0.4412	0.5675	0.5675	0.5674
mgh14	7.537	7.262	7.466	7.318	6.657	7.245
mgh17	$3.857 \cdot 10^{68}$	1.374	$8.32 \cdot 10^{46}$	$1.135 \cdot 10^{70}$	0.3561	$2.527 \cdot 10^{48}$
mgh20	1.321	1.262	1.279	1.487	1.427	1.445
mgh21	$3.941 \cdot 10^1$	$1.556 \cdot 10^1$	$2.496 \cdot 10^1$	$1.758 \cdot 10^1$	3.972	$1.083 \cdot 10^1$
mgh22	2.184	2.183	2.183	1.219	1.219	1.219
mgh25	$1.647 \cdot 10^1$	$1.647 \cdot 10^1$	$1.643 \cdot 10^1$	$5.02 \cdot 10^1$	$5.021 \cdot 10^1$	$5.0 \cdot 10^1$
mgh27	0.4891	0.4876	0.488	0.7019	0.6896	0.6995
mgh30	1.408	1.382	1.402	0.7588	0.7236	0.755
mgh31	2.063	1.846	2.01	1.217	0.8228	1.166
mgh33	$4.516 \cdot 10^3$	$4.516 \cdot 10^3$	$4.506 \cdot 10^3$	$6.392 \cdot 10^3$	$6.392 \cdot 10^3$	$6.381 \cdot 10^3$
mgh34	$2.378 \cdot 10^3$	$2.378 \cdot 10^3$	$2.372 \cdot 10^3$	$3.207 \cdot 10^3$	$3.207 \cdot 10^3$	$3.204 \cdot 10^3$
mgh35	$1.221 \cdot 10^3$	$3.846 \cdot 10^3$	$3.717 \cdot 10^2$	$2.563 \cdot 10^4$	$4.247 \cdot 10^4$	$7.365 \cdot 10^3$

problem a stationary point x_N^* of the nominal problem (7.4) and $Z_\epsilon(x_N^*)$ defined by

$$Z_\epsilon(x_N^*) = \mathbb{E}_{N(0,\epsilon I)}[X(x_N^*)] + \mathrm{StD}_{N(0,\epsilon I)}[X(x_N^*)], \quad X(x_N^*)(\xi) = \frac{f_0(x_N^* + \xi) - f_0(x_N^*)}{\max\{1, |f_0(x_N^*)|\}}$$

Table 5

Median number of objective function, gradient, and Hessian evaluations required by `Ipopt` *for the nominal (N), distributionally robust (DR), and stochastic optimization problem (S) of all selected test problems. The number of evaluations for the approximate DROPs* (7.2) *are the sum of all evaluations used within Algorithm* 3.1.

	N	DR	S	N	DR	S	N	S
ϵ	#-f_0	#-$\widetilde{F}_0$	#-f_0	#-∇f_0	#-$\nabla_x \widetilde{F}_0$	#-∇f_0	#-$\nabla^2 f_0$	#-$\nabla^2 f_0$
10^{-3}	14	120	$1.25 \cdot 10^4$	14	37.5	$1.2 \cdot 10^4$	13	$1.1 \cdot 10^4$
10^{-2}	14	190.5	$1.0 \cdot 10^4$	14	56	$1.0 \cdot 10^4$	13	$0.9 \cdot 10^3$

Table 6

$Z_\epsilon(x_N^*)$ and number of parameters p of problems from the Moré–Garbow–Hillstrom test set with $Z_\epsilon(x_N^*)$ exceeding 10^{-1}, where $\epsilon = 10^{-3}$.

Problem	p	$Z_\epsilon(x_N^*)$	Problem	p	$Z_\epsilon(x_N^*)$	Problem	p	$Z_\epsilon(x_N^*)$
mgh01	2	0.5778	mgh13	4	0.1262	mgh27	10	0.1403
mgh03	2	$1.006 \cdot 10^7$	mgh14	4	1.911	mgh30	10	0.2737
mgh04	2	$1.234 \cdot 10^9$	mgh17	5	$8.83 \cdot 10^{24}$	mgh31	10	0.3551
mgh06	2	1.903	mgh20	6	0.3353	mgh33	10	$5.505 \cdot 10^2$
mgh07	3	0.4727	mgh21	20	7.056	mgh34	10	$2.243 \cdot 10^2$
mgh10	3	$3.524 \cdot 10^9$	mgh22	20	0.4509	mgh35	10	0.8223
mgh11	3	$2.567 \cdot 10^{127}$	mgh25	10	1.282			

and selected all problems fulfilling $Z_\epsilon(x_N^*) \geq 10^{-1}$ for $\epsilon = 10^{-3}$; see Table 6.

A related approach has been used in [4] to investigate uncertain linear programs.

Acknowledgments. We would like to thank the anonymous referees and the associate editor for their valuable comments and suggestions.

REFERENCES

[1] R. Andreani, G. Haeser, and J. M. Martínez, *On sequential optimality conditions for smooth constrained optimization*, Optimization, 60 (2011), pp. 627–641, https://doi.org/10.1080/02331930903578700.

[2] A. Ben-Tal, D. den Hertog, and J.-P. Vial, *Deriving robust counterparts of nonlinear uncertain inequalities*, Math. Program., 149 (2015), pp. 265–299, https://doi.org/10.1007/s10107-014-0750-8.

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

[4] A. Ben-Tal and A. Nemirovski, *Robust solutions of linear programming problems contaminated with uncertain data*, Math. Program., 88 (2000), pp. 411–424, https://doi.org/10.1007/PL00011380.

[5] A. Ben-Tal and A. Nemirovski, *Lectures on Modern Convex Optimization*, MPS-SIAM Ser. Optim. 2, SIAM, Philadelphia, 2001, https://doi.org/10.1137/1.9780898718829.

[6] A. Ben-Tal and M. Teboulle, *Hidden convexity in some nonconvex quadratically constrained quadratic programming*, Math. Program., 72 (1996), pp. 51–63, https://doi.org/10.1007/BF02592331.

[7] D. Bertsimas, D. B. Brown, and C. Caramanis, *Theory and applications of robust optimization*, SIAM Rev., 53 (2011), pp. 464–501, https://doi.org/10.1137/080734510.

[8] D. Bertsimas, O. Nohadani, and K. M. Teo, *Nonconvex robust optimization for problems with constraints*, INFORMS J. Comput., 22 (2010), pp. 44–58, https://doi.org/10.1287/ijoc.1090.0319.

[9] D. Bertsimas, O. Nohadani, and K. M. Teo, *Robust optimization for unconstrained simulation-based problems*, Oper. Res., 58 (2010), pp. 161–178, https://doi.org/10.1287/opre.1090.0715.

[10] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A fresh approach to numerical computing*, SIAM Rev., 59 (2017), pp. 65–98, https://doi.org/10.1137/141000671.

[11] J. F. Bonnans and A. Shapiro, *Perturbation Analysis of Optimization Problems*, Springer Ser. Oper. Res., Springer Science & Business Media, New York, 2000, https://doi.org/10.1007/978-1-4612-1394-9.

[12] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004, https://doi.org/10.1017/CBO9780511804441.

[13] V. V. Buldygin and Y. V. Kozachenko, *Metric Characterization of Random Variables and Random Processes*, Transl. Math. Monogr. 188, American Mathematical Society, Providence, RI, 2000.

[14] J. V. Burke and T. Hoheisel, *Epi-convergent smoothing with applications to convex composite functions*, SIAM J. Optim., 23 (2013), pp. 1457–1479, https://doi.org/10.1137/120889812.

[15] J. V. Burke and T. Hoheisel, *Epi-convergence properties of smoothing by infimal convolution*, Set-Valued Var. Anal., 25 (2017), pp. 1–23, https://doi.org/10.1007/s11228-016-0362-y.

[16] J. V. Burke, T. Hoheisel, and C. Kanzow, *Gradient consistency for integral-convolution smoothing functions*, Set-Valued Var. Anal., 21 (2013), pp. 359–376, https://doi.org/10.1007/s11228-013-0235-6.

[17] J. V. Burke, A. S. Lewis, and M. L. Overton, *A robust gradient sampling algorithm for nonsmooth, nonconvex optimization*, SIAM J. Optim., 15 (2005), pp. 751–779, https://doi.org/10.1137/030601296.

[18] X. Chen, *Smoothing methods for nonsmooth, nonconvex minimization*, Math. Program., 134 (2012), pp. 71–99, https://doi.org/10.1007/s10107-012-0569-0.

[19] X. Chen, H. Qi, L. Qi, and K.-L. Teo, *Smooth convex approximation to the maximum eigenvalue function*, J. Global Optim., 30 (2004), pp. 253–270, https://doi.org/10.1007/s10898-004-8271-2.

[20] Z. Chen, M. Sim, and H. Xu, *Distributionally robust optimization with infinitely constrained ambiguity sets*, Oper. Res., 67 (2019), pp. 1328–1344, https://doi.org/10.1287/opre.2018.1799.

[21] F. H. Clarke, *Generalized gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262, https://doi.org/10.1090/S0002-9947-1975-0367131-6.

[22] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990, https://doi.org/10.1137/1.9781611971309.

[23] E. Delage and Y. Ye, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Oper. Res., 58 (2010), pp. 595–612, https://doi.org/10.1287/opre.1090.0741.

[24] M. Diehl, H. G. Bock, and E. Kostina, *An approximation technique for robust nonlinear optimization*, Math. Program., 107 (2006), pp. 213–230, https://doi.org/10.1007/s10107-005-0685-1.

[25] J. Dutta, K. Deb, R. Tulshyan, and R. Arora, *Approximate KKT points and a proximity measure for termination*, J. Global Optim., 56 (2013), pp. 1463–1499, https://doi.org/10.1007/s10898-012-9920-5.

[26] P. M. Esfahani and D. Kuhn, *Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations*, Math. Program., 171 (2018), pp. 115–166, https://doi.org/10.1007/s10107-017-1172-1.

[27] A. V. Fiacco and Y. Ishizuka, *Sensitivity and stability analysis for nonlinear programming*, Ann. Oper. Res., 27 (1990), pp. 215–235, https://doi.org/10.1007/BF02055196.

[28] J. Fiala, M. Kočvara, and M. Stingl, *PENLAB: A MATLAB Solver for Nonlinear Semidefinite Optimization*, preprint, https://arxiv.org/abs/1311.5240, 2013.

[29] A. Forsgren, P. E. Gill, and M. H. Wright, *Interior methods for nonlinear optimization*, SIAM Rev., 44 (2002), pp. 525–597, https://doi.org/10.1137/S0036144502414942.

[30] R. Gao and A. J. Kleywegt, *Distributionally Robust Stochastic Optimization with Wasserstein Distance*, preprint, https://arxiv.org/abs/1604.02199, 2016.

[31] J. Goh and M. Sim, *Distributionally robust optimization and its tractable approximations*, Oper. Res., 58 (2010), pp. 902–917, https://doi.org/10.1287/opre.1090.0795.

[32] W. W. Hogan, *Point-to-set maps in mathematical programming*, SIAM Rev., 15 (1973), pp. 591–603, https://doi.org/10.1137/1015073.

[33] R. A. Horn and C. R. Johnson, *Matrix Analysis*, 2nd ed., Cambridge University Press, Cambridge, 2013, https://doi.org/10.1017/9781139020411.

[34] B. Houska and M. Diehl, *Nonlinear robust optimization via sequential convex bilevel programming*, Math. Program., 142 (2013), pp. 539–577, https://doi.org/10.1007/s10107-012-0591-2.

[35] C. Kanzow and A. Schwartz, *The price of inexactness: Convergence properties of relaxation methods for mathematical programs with complementarity constraints revisited*, Math. Oper. Res., 40 (2014), pp. 253–275, https://doi.org/10.1287/moor.2014.0667.

[36] N. Karmitsa, *Solver-o-matic*, http://napsu.karmitsa.fi/solveromatic/, 2012, accessed September 12, 2019.
[37] N. Karmitsa, A. Bagirov, and M. M. Mäkelä, *Comparing different nonsmooth minimization methods and software*, Optim. Methods Softw., 27 (2012), pp. 131–153, https://doi.org/10.1080/10556788.2010.526116.
[38] P. Kolvenbach, O. Lass, and S. Ulbrich, *An approach for robust PDE-constrained optimization with application to shape optimization of electrical engines and of dynamic elastic structures under uncertainty*, Optim. Eng., 19 (2018), pp. 697–731, https://doi.org/10.1007/s11081-018-9388-3.
[39] O. Lass and S. Ulbrich, *Model order reduction techniques with a posteriori error control for nonlinear robust optimization governed by partial differential equations*, SIAM J. Sci. Comput., 39 (2017), pp. S112–S139, https://doi.org/10.1137/16M108269X.
[40] A. S. Lewis, *Derivatives of spectral functions*, Math. Oper. Res., 21 (1996), pp. 576–588, https://doi.org/10.1287/moor.21.3.576.
[41] A. S. Lewis, *Nonsmooth analysis of eigenvalues*, Math. Program., 84 (1999), pp. 1–24, https://doi.org/10.1007/s10107980004a.
[42] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Math. Program., 141 (2013), pp. 135–163, https://doi.org/10.1007/s10107-012-0514-2.
[43] M. M. Mäkelä, *Multiobjective Proximal Bundle Method for Nonconvex Nonsmooth Optimization: Fortran Subroutine MPBNGC* 2.0, Reports of the Department of Mathematical Information Technology, Series B, Scientific Computing B 13/2003, University of Jyväskylä, Jyväskylä, 2003.
[44] M. M. Mäkelä, N. Karmitsa, and O. Wilppu, *Proximal bundle method for nonsmooth and nonconvex multiobjective optimization*, in Mathematical Modeling and Optimization of Complex Structures, P. Neittaanmäki, S. Repin, and T. Tuovinen, eds., Comput. Methods Appl. Sci. 40, Springer, Cham, 2016, pp. 191–204, https://doi.org/10.1007/978-3-319-23564-6_12.
[45] M. M. Mäkelä and P. Neittaanmäki, *Nonsmooth Optimization: Analysis and Algorithms with Applications to Optimal Control*, World Scientific Singapore, Singapore, 1992, https://doi.org/10.1142/1493.
[46] J. J. Moré, B. S. Garbow, and K. E. Hillstrom, *Testing unconstrained optimization software*, ACM Trans. Math. Softw., 7 (1981), pp. 17–41.
[47] J. J. Moré and D. C. Sorensen, *Computing a trust region step*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 553–572, https://doi.org/10.1137/0904038.
[48] Yu. Nesterov, *Smoothing technique and its applications in semidefinite optimization*, Math. Program., 110 (2007), pp. 245–259, https://doi.org/10.1007/s10107-006-0001-8.
[49] Yu. Nesterov and A. Nemirovski, *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM Stud. Appl. Math. 13, SIAM, Philadelphia, 1994, https://doi.org/10.1137/1.9781611970791.
[50] B. O'Donoghue, E. Chu, N. Parikh, and S. Boyd, *Conic optimization via operator splitting and homogeneous self-dual embedding*, J. Optim. Theory Appl., 169 (2016), pp. 1042–1068, https://doi.org/10.1007/s10957-016-0892-3.
[51] E. Polak, J. O. Royset, and R. Womersley, *Algorithms with adaptive smoothing for finite minimax problems*, J. Optim. Theory Appl., 119 (2003), pp. 459–484, https://doi.org/10.1023/B:JOTA.0000006685.60019.3e.
[52] J. Revels, M. Lubin, and T. Papamarkou, *Forward-Mode Automatic Differentiation in Julia*, preprint, https://arxiv.org/abs/1607.07892, 2016.
[53] A. Shapiro, *Distributionally robust stochastic programming*, SIAM J. Optim., 27 (2017), pp. 2258–2275, https://doi.org/10.1137/16M1058297.
[54] A. Shapiro and A. Kleywegt, *Minimax analysis of stochastic problems*, Optim. Methods Softw., 17 (2002), pp. 523–542, https://doi.org/10.1080/1055678021000034008.
[55] A. M.-C. So, *Moment inequalities for sums of random matrices and their applications in optimization*, Math. Program., 130 (2011), pp. 125–151, https://doi.org/10.1007/s10107-009-0330-5.
[56] D. C. Sorensen, *Newton's method with a model trust region modification*, SIAM J. Numer. Anal., 19 (1982), pp. 409–426, https://doi.org/10.1137/0719026.
[57] S. Steffensen and M. Ulbrich, *A new relaxation scheme for mathematical programs with equilibrium constraints*, SIAM J. Optim., 20 (2010), pp. 2504–2539, https://doi.org/10.1137/090748883.
[58] R. J. Stern and H. Wolkowicz, *Indefinite trust region subproblems and nonsymmetric eigenvalue perturbations*, SIAM J. Optim., 5 (1995), pp. 286–313, https://doi.org/10.1137/0805016.

[59] P. D. Tao and L. T. H. An, *Lagrangian stability and global optimality in nonconvex quadratic minimization over Euclidean balls and spheres*, J. Convex Anal., 2 (1995), pp. 263–276.

[60] N.-K. Tsing, M. K. Fan, and E. I. Verriest, *On analyticity of functions involving eigenvalues*, Linear Algebra Appl., 207 (1994), pp. 159–180, https://doi.org/10.1016/0024-3795(94)90009-4.

[61] A. Wächter and L. T. Biegler, *On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming*, Math. Program., 106 (2006), pp. 25–57, https://doi.org/10.1007/s10107-004-0559-y.

[62] M. J. Weinstein and A. V. Rao, *Algorithm* 984: *ADiGator, a toolbox for the algorithmic differentiation of mathematical functions in MATLAB using source transformation via operator overloading*, ACM Trans. Math. Softw., 44 (2017), 21, https://doi.org/10.1145/3104990.

[63] W. Wiesemann, D. Kuhn, and M. Sim, *Distributionally robust convex optimization*, Oper. Res., 62 (2014), pp. 1358–1376, https://doi.org/10.1287/opre.2014.1314.

[64] M. Xu, J. J. Ye, and L. Zhang, *Smoothing SQP methods for solving degenerate nonsmooth constrained optimization problems with applications to bilevel programs*, SIAM J. Optim., 25 (2015), pp. 1388–1410, https://doi.org/10.1137/140971580.

[65] Y. Xu, W. Sun, and L. Qi, *A feasible direction method for the semidefinite program with box constraints*, Appl. Math. Lett., 24 (2011), pp. 1874–1881, https://doi.org/10.1016/j.aml.2011.05.010.

[66] H. Yamashita and H. Yabe, *A survey of numerical methods for nonlinear semidefinite programming*, J. Oper. Res. Soc. Japan, 58 (2015), pp. 24–60, https://doi.org/10.15807/jorsj.58.24.

[67] Y. Zhang, *General robust-optimization formulation for nonlinear programming*, J. Optim. Theory Appl., 132 (2007), pp. 111–124, https://doi.org/10.1007/s10957-006-9082-z.

[68] C. Zhao and Y. Guan, *Data-driven risk-averse stochastic optimization with Wasserstein metric*, Oper. Res. Lett., 46 (2018), pp. 262–267, https://doi.org/10.1016/j.orl.2018.01.011.