Noname manuscript No.
(will be inserted by the editor)

Derivative-Free Robust Optimization by Outer Approximations

Matt Menickelly · Stefan M. Wild

the date of receipt and acceptance should be inserted later

Abstract We develop an algorithm for minimax problems that arise in robust optimization in the absence of objective function derivatives. The algorithm utilizes an extension of methods for inexact outer approximation in sampling a potentially infinite-cardinality uncertainty set. Clarke stationarity of the algorithm output is established alongside desirable features of the model-based trust-region subproblems encountered. We demonstrate the practical benefits of the algorithm on a new class of test problems.

Keywords Derivative-Free Optimization · Robust Optimization · Outer Approximation Algorithms · Manifold Sampling

Mathematics Subject Classification (2010) 90C56 · 65K10 · 90C47 · 90C30

1 Introduction

We consider the *robust optimization problem*

$$\min_{x\in\mathbb{R}^n} \max_{u\in\mathcal{U}} f(x,u), \tag{1}$$

where $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ and $\mathcal{U} \subset \mathbb{R}^m$ is called the *uncertainty set.* The minimax problem (1) can represent the minimization of the worst-case objective function under deterministic uncertainty in the problem data $u \in \mathcal{U}$ [2, 4].

For any subset $\hat{\mathcal{U}} \subseteq \mathcal{U}$ we define

$$\Psi_{\hat{\mathcal{U}}}(x) \triangleq \max_{u\in\hat{\mathcal{U}}} f(x,u)$$

and use the shorthand $\Psi \triangleq \Psi_{\mathcal{U}}$. Consequently, an equivalent representation of (1) is the implicitly robustified form

$$\min_{x\in\mathbb{R}^n} \Psi(x),$$

Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL 60439, USA
E-mail: mmenickelly@anl.gov; E-mail: wild@anl.gov

which we note is not equivalent to the *nominal optimization problem*

$$\min_{x\in\mathbb{R}^n} f(x,\hat{u}),$$

where $\hat{u}\in\mathcal{U}$ represents a nominal data value.

In this paper, we address the case where derivatives of the function f in (1) (and hence the function Ψ) are unavailable or inaccurate. Such situations arise, for example, when only a deterministic black-box ("zero-order") oracle that computes values of $f(x,u)$ is available [3, 7, 10, 12, 13].

Central to our algorithmic development is an optimality measure for (1) and its tractable approximation by using a finitely generated set. Section 2 states properties of this optimality measure, which is then used in Section 3 in an inexact method of outer approximations as obtained by iteratively tightened relaxations of the uncertainty set $\mathcal{U}$ [22]. This optimality measure relies on access to values of the derivatives ∇f, which we relax in Section 4 by using model-based approximations that can be constructed solely from values of the function f. The resulting algorithm employs an iterative, model-building framework similar to that employed by the manifold sampling of [19, 20].

In Section 5, we analyze the algorithm and establish Clarke stationarity of the limit points obtained. Theoretical algorithms for robust optimization are known to suffer from undesirable computational requirements [1, 5, 6, 8, 14, 17, 23]. We address practical considerations for implementations of the proposed algorithm in Section 6. Our implementation is then tested in Section 7, and we highlight some of the additional expense incurred when operating without derivative values.

2 Optimality Measure

To set the stage, we state several assumptions concerning (1) that will be used throughout and that mirror those in [22, Assumption 3.4.1].

Assumption 1 *The following hold.*

a. ***(Local Lipschitz continuity of f and $\nabla_x f$ everywhere)*** *The function $f(\cdot,\cdot)$ and, for any $u\in\mathcal{U}$, its partial gradient $\nabla_x f(\cdot,u)$ are Lipschitz continuous over any bounded subset of $\mathbb{R}^n\times\mathbb{R}^m$ and $\mathbb{R}^n$, respectively.*
b. ***(Compactness of $\mathcal{U}$)*** *$\mathcal{U}\subset\mathbb{R}^m$ is a compact set.*
c. ***(Solution existence)*** *There exists a (finite) solution to (1).*

We remark that Assumption 1.a relies on the derivative $\nabla_x f$, which we assume exists but do not assume to be available for use by an algorithm for solving (1). The uncertainty set $\mathcal{U}$ is a modeling choice, and thus Assumption 1.b is largely nonrestrictive. Assumption 1.c may be difficult to verify a priori.

We first consider the following optimality measure for (1):

$$\begin{aligned}\Theta(x) &\triangleq \min_{h\in\mathbb{R}^n}\max_{u\in\mathcal{U}}\left\{f(x,u)+\langle\nabla_x f(x,u),h\rangle+\frac{1}{2}\|h\|^2\right\}-\Psi(x)\\ &=\min_{h\in\mathbb{R}^n}\theta(x,h),\end{aligned}\tag{2}$$

where

$$\theta(x,h)\triangleq\max_{u\in\mathcal{U}}\left\{f(x,u)+\langle\nabla_x f(x,u),h\rangle+\frac{1}{2}\|h\|^2\right\}-\Psi(x).\tag{3}$$

For a fixed u, the objective in the minimax problem in (2) corresponds to a second-order convex approximation of $f(\cdot, u)$ at x. Thus, the minimax problem in (2) represents a minimization over the upper envelope (described by $\mathcal{U}$) of a family of convex quadratics. The final term, $-\Psi(x)$, can be viewed as shifting the optimality measure (2). Since

$$\min_{h\in\mathbb{R}^n} \theta(x,h) \leq \theta(x,\mathbf{0}) = \max_{u\in\mathcal{U}} f(x,u) - \Psi(x) = 0,$$

we have the following proposition.

Proposition 1 *Let Assumption 1 hold; then, for all $x \in \mathbb{R}^n$, $\Theta(x) \leq 0$.*

We now state additional properties of the optimality measure $\Theta(x)$; proofs can be found in Appendix A.

For any $\hat{\mathcal{U}} \subseteq \mathcal{U}$, we define the set

$$\mathcal{D}_{f,\hat{\mathcal{U}}}(x) \triangleq \mathbf{co}\left\{ \begin{bmatrix} \Psi_{\hat{\mathcal{U}}}(x) - f(x,u) \\ \nabla_x f(x,u) \end{bmatrix} : u \in \hat{\mathcal{U}} \right\}, \tag{4}$$

where **co** denotes the convex hull. We use $\xi_0 \in \mathbb{R}$ and $\xi \in \mathbb{R}^n$ to denote a generic element (ξ_0, ξ) of the set in (4).

The following proposition establishes an equivalent definition of $\Theta(x)$, which will be interpreted as the value of a dual problem used by our algorithm.

Proposition 2 *Let Assumption 1 hold; then, for all $x \in \mathbb{R}^n$,*

$$\Theta(x) = -\min_{\xi_0,\xi} \left\{ \xi_0 + \frac{1}{2}\|\xi\|^2 : (\xi_0,\xi) \in \mathcal{D}_{f,\mathcal{U}}(x) \right\}. \tag{5}$$

The next proposition establishes the biconditional relationship between the optimality measure Θ and Clarke stationarity of Ψ.

Proposition 3 *Let Assumption 1 hold; then, for all $x \in \mathbb{R}^n$, $\mathbf{0} \in \partial\Psi(x)$ if and only if $\Theta(x) = 0$.*

We rely on the following auxiliary result to prove convergence of our algorithm.

Proposition 4 *Let Assumption 1 hold; then $\Theta(x)$ is a continuous function of x.*

Having defined an optimality measure for (1) with its equivalent (dual) measure (5), we also define, for $\hat{\mathcal{U}} \subseteq \mathcal{U}$, the *approximate optimality measure*

$$\Theta_{\hat{\mathcal{U}}}(x) \triangleq -\min_{\xi_0,\xi} \left\{ \xi_0 + \frac{1}{2}\|\xi\|^2 : (\xi_0,\xi) \in \mathcal{D}_{f,\hat{\mathcal{U}}}(x) \right\}. \tag{6}$$

Algorithm 1: Inexact Method of Outer Approximations

```
1 Choose $\left\{\left(\epsilon_k, \varOmega^k\right)\right\}_{k=0}^{\infty}$ satisfying Assumption 2 and choose starting point $x^0 \in \mathbb{R}^n$.
2 Compute any $u^1 \in \operatorname*{argmax}_{u \in \varOmega^0} f(x^0, u)$ and set $\mathfrak{U}^0 \leftarrow \{u^1\}$.
3 Set $k \leftarrow 0$.
4 while true do
5     Compute any $x^{k+1}$ such that $\varTheta_{\mathfrak{U}^k}(x^{k+1}) \geq -\epsilon_k$.
6     Compute any $u' \in \operatorname*{argmax}_{u \in \varOmega^k} f(x^{k+1}, u)$.
7     Augment $\mathfrak{U}^{k+1} \leftarrow \mathfrak{U}^k \cup \{u'\}$.
8     $k \leftarrow k + 1$.
9 end
```

3 Inexact Method of Outer Approximations

Our approach is based on the method of outer approximations [22, Section 3.4.5], which is a type of cutting-plane method (see, e.g., [9, 15, 16, 18]). [1] An inexact method of outer approximation does not require exact solutions of the alternating block subproblems

$$\left(\min_{x \in \mathbb{R}^n} \varPsi_{\hat{\mathcal{U}}}(x), \qquad \max_{u \in \mathcal{U}} f(\hat{x}, u) \right).$$

In Algorithm 1 we state the inexact method from [22, Algorithm 3.4.26]. The algorithm we propose in the next section can be viewed as a derivative-free extension of Algorithm 1. We note that an exact method of outer approximation may be obtained by setting $\left(\epsilon_k = 0, \varOmega^k = \mathcal{U}\right)$ for all k.

Algorithm 1 entails the iterative solution of alternating block subproblems: Line 5 is an ϵ_k-accurate unconstrained minimization subproblem over the variables x, while the subproblem in Line 6 is an $\varOmega^k$-constrained maximization over the variables u. To prove a convergence result for Algorithm 1, we require the following assumptions on the sequence $\left\{\left(\epsilon_k, \varOmega^k\right)\right\}_{k=0}^{\infty}$.

Assumption 2 *The following properties hold for the family of subsets* $\{\varOmega^k\}_{k=0}^{\infty}$ *and the tolerances* $\{\epsilon_k\}_{k=0}^{\infty}$*:*

a. $\varOmega^k \subseteq \mathcal{U}$ *for all* $k = 0, 1, \ldots$.
b. For all $k = 0, 1, \ldots$ *and all* $\hat{x} \in \mathbb{R}^n$*, the subproblem* $\max_{u \in \varOmega^k} f(\hat{x}, u)$ *can be solved exactly.*
c. $\min_{x \in \mathbb{R}^n} \max_{u \in \varOmega^k} f(x, u)$ *has a solution for all* $k = 0, 1, \ldots$.
d. There exists a strictly positive, monotone decreasing function $\delta : \mathbb{N} \to \mathbb{R}$ *satisfying* $\lim_{k\to\infty} \delta(k) = 0$ *and a constant* $\kappa_0 > 0$ *such that for all* $u \in \mathcal{U}$*, there exists* $u' \in \varOmega^k$ *such that* $\|u - u'\| \leq \kappa_0 \delta(k)$ *for all* $k = 0, 1, \ldots$.
e. For all $k = 0, 1, \ldots$, $\epsilon_k \in [0, 1]$ *and* $\lim_{k\to\infty} \epsilon_k = 0$.

We note that Assumption 2.b is implicitly a statement on tractability and can be satisfied when, for instance, $|\varOmega^k| < \infty$ for all $k = 0, 1, \ldots$. We also note that

[1] Our selection of such an approach is informed by a recent study [5] highlighting merits of cutting plane methods in various robust optimization settings.

Assumption 2.c ensures that an analog of Assumption 1.c holds when $\mathcal{U}$ is replaced by Ω.

Assumption 2.d is discouraging in that it effectively requires the (potentially finite) sample Ω^k to be asymptotically dense in $\mathcal{U}$. This requirement is inevitable without further assumptions on f, however, since Line 6 of Algorithm 1 seeks an approximation to the *global* maximizer over $\mathcal{U}$ of $f(x^{k+1}, u)$ in order to recover the convergence results of the exact method of outer approximation [22]. Assumption 2.d is also the assumption that inspires us to separate the roles of Ω^k and $\mathfrak{U}^k$; the latter forms the basis for our approximate optimality measure but need not cover $\mathcal{U}$ asymptotically.

Because the convergence of our proposed algorithm, Algorithm 2, depends on the convergence of Algorithm 1, we prove the following theorem.

Theorem 1 *Suppose Assumptions 1 and 2 hold. Let x^* be an accumulation point of the sequence $\{x^k\}_{k=1}^\infty$ generated by Algorithm 1. Then $\Theta(x^*) = 0$; thus, by Proposition 3, $\mathbf{0} \in \partial\Psi(x^*)$.*

The proof, provided in Appendix B, shows that as Algorithm 1 progresses,

1. the sequence of finite max functions $\Psi_{\Omega^k}(x)$ are, uniformly over $x \in \mathbb{R}^n$, arbitrarily good approximations of $\Psi(x)$ and
2. the sequence of optimality measures $\Theta_{\Omega^k}(x)$ are, uniformly over $x \in \mathbb{R}^n$, arbitrarily good approximations of the optimality measure $\Theta(x)$.

4 A New Derivative-Free Algorithm

The main challenge of creating a derivative-free extension of Algorithm 1 is in computing sufficiently accurate approximations to $\Theta_{\mathfrak{U}^k}(x^k)$ in the absence of ∇f values. We show that by computing a parsimonious subset of sufficiently accurate gradient approximations, as is done in the framework of manifold sampling [19, 20], we are able to compute such approximations of $\Theta_{\mathfrak{U}^k}(x^k)$. In particular, by using a manifold sampling framework, we do not need to maintain gradient approximations corresponding to every element in $\mathfrak{U}^k$ from Algorithm 1, potentially yielding savings on the computational burden of function evaluations in a derivative-free setting.

To develop a derivative-free variant of Algorithm 1, for which we proved convergence in Theorem 1, our primary concern is in developing a suitable approximation of Line 5 in Algorithm 1. We begin by introducing notation specific to our lack of derivative information.

For the inner method used to approximate Line 5 in Algorithm 1, we index iterations by t, while we continue to index the outer iterations of Algorithm 1 by k. Moreover, to denote primal iterates within the inner method, we use y^t instead of x^k, the latter being reserved for iterates of the outer method.

For a fixed $u^j \in \mathfrak{U}^k$, where $j \in \{1, \ldots, |\mathfrak{U}^k|\}$, we denote the jth local model in the tth inner iteration by $m_j^t : \mathbb{R}^n \to \mathbb{R}$. Within the algorithm, we enforce that $m_j^t(y)$ is a *fully linear* (see, e.g., [12]) model of $f(y, u^j)$ for all $y \in \mathcal{B}(y^t, \Delta_t)$, where $y^t \in \mathbb{R}^n$ is a primal iterate and $\Delta_t > 0$ is a trust-region radius.[2]

[2] We note that the proposed algorithm and its analysis could also employ inexact gradient values, provided that these gradients satisfy the approximation specified in Assumption 3.

Assumption 3 *Given $\Delta > 0$, $y^t \in \mathbb{R}^n$, and $\mathfrak{U}^k \subseteq \mathcal{U}$, there exist constants $\kappa_f > 0$ and $\kappa_g > 0$, independent of Δ, y^t, and $\mathfrak{U}^k$, such that for all $u^j \in \mathfrak{U}^k$*

$$\begin{aligned} |f(y^t+s,u^j) - m_j^t(y^t+s)| &\le \kappa_f \Delta^2 \qquad \forall s \in \mathcal{B}(0,\Delta) \\ \|\nabla f(y^t+s,u^j) - \nabla m_j^t(y^t+s)\| &\le \kappa_g \Delta \qquad \forall s \in \mathcal{B}(0,\Delta). \end{aligned}$$

Additionally, the interpolation condition $m_j^t(y^t) = f(y^t,u^j)$ holds for all $j = 1,\dots,|\mathfrak{U}^k|$.

As in a manifold sampling algorithm, we employ in iteration t a *set of generator indices*, $J^{t,k}$, corresponding to elements $\{u^j \in \mathfrak{U}^k : j \in J^{t,k}\}$. At the end of this section we discuss how $J^{t,k} \subseteq \{1,\dots,|\mathfrak{U}^k|\}$ is selected in iteration t. Given $J^{t,k}$, we define a matrix of model gradients and vector of function values, respectively, by

$$\begin{aligned} G^t &\triangleq \left[\nabla m^t_{\sigma(1)}(y^t),\dots,\nabla m^t_{\sigma(|J^{t,k}|)}(y^t)\right] \in \mathbb{R}^{n\times|J^{t,k}|} \\ F^t &\triangleq \left[f(y^t,u^{\sigma(1)}),\dots,f(y^t,u^{\sigma(|J^{t,k}|)})\right]^\top \in \mathbb{R}^{|J^{t,k}|}, \end{aligned}$$

where $\sigma : \{1,\dots,|\mathfrak{U}^k|\} \mapsto \{1,\dots,|\mathfrak{U}^k|\}$ is a permutation such that $\cup_{i=1,\dots,|J^{t,k}|}\sigma(i) = J^{t,k}$.

We now define a natural approximation $\tilde{\Theta}_{\mathfrak{U}^k}$ to the approximation $\Theta_{\mathfrak{U}^k}$ of Θ. To this end, we define a set analogous to that in (4):

$$\mathcal{D}_{m^t,\mathfrak{U}^k}(y^t) \triangleq \mathbf{co}\left\{\begin{bmatrix}\Psi_{\mathfrak{U}^k}(y^t) - m_j^t(y^t) \\ \nabla_x m_j^t(y^t)\end{bmatrix} : u^j \in \mathfrak{U}^k\right\}. \tag{7}$$

Analogously to how we let $(\xi_0,\xi) \in \mathbb{R}\times\mathbb{R}^n$ denote elements of $\mathcal{D}_{f,\mathcal{U}}(x)$ in (5), we employ arbitrary elements $(z_0,z) \in \mathbb{R}\times\mathbb{R}^n$ of $\mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$ in defining the approximate inexact measure

$$\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t) \triangleq -\min_{z_0,z}\left\{z_0 + \frac{1}{2}\|z\|^2 : (z_0,z) \in \mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)\right\}. \tag{8}$$

We remark that while (8) considers the convex set $\mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$ built on information from all $|\mathfrak{U}^k|$ models, we generally will not need to construct this many models, as will be elucidated in our algorithm; in fact, we need to construct and ensure the fully linear approximation of only $|J^{t,k}|$ many models in inner iteration t.

The following lemma bounds the error made by $\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t)$ in approximating $\Theta_{\mathfrak{U}^k}(y^t)$.

Lemma 1 *Let Assumption 3 hold. For all $(z_0,z) \in \mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$, there exists $(\xi_0(z_0,z), \xi(z_0,z)) \in \mathcal{D}_{f,\mathcal{U}}(y^t)$ so that*

$$\begin{aligned} z_0 &= \xi_0(z_0,z) \\ \|z - \xi(z_0,z)\| &\le \kappa_g \Delta_k. \end{aligned}$$

Proof Since $\mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$ is convex and finitely generated (hence, compact), Caratheodory's theorem ensures that any $(z_0,z) \in \mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$ can be expressed as a positive convex combination of $N \le n+1$ of its generators. Without loss of generality, let these generators correspond to the first N elements of $\mathfrak{U}^k$. That is, there exist $\lambda_1, \dots, \lambda_N \in (0,1]$ with $\sum_{j=1}^N \lambda_j = 1$ such that

$$\begin{bmatrix} z_0 \\ z \end{bmatrix} = \sum_{j=1}^{N} \lambda_j \begin{bmatrix} \Psi_{\mathfrak{U}^k}(y^t) - m_j^t(y^t, u^j) \\ \nabla_x m_j^t(y^t) \end{bmatrix}. \tag{9}$$

By using the same λ_j as in (9), we let

$$\begin{bmatrix} \xi_0(z_0, z) \\ \xi(z_0, z) \end{bmatrix} = \sum_{j=1}^{N} \lambda_j \begin{bmatrix} \Psi_{\mathfrak{U}^k}(y^t) - f(y^t, u^j) \\ \nabla_x f(y^t, u^j) \end{bmatrix}$$

and note that $|z_0 - \xi_0(z_0, z)| = 0$.

By the definition in (4),

$$\begin{bmatrix} \Psi_{\mathfrak{U}^k}(y^t) - f(y^t, u^j) \\ \nabla_x f(y^t, u^j) \end{bmatrix} \in \mathcal{D}_{f,\mathcal{U}}(y^t), \qquad j = 1, \dots, N,$$

and thus since $\mathcal{D}_{f,\mathcal{U}}(y^t)$ is a convex set, we have that $(\xi_0(z_0, z), \xi(z_0, z)) \in \mathcal{D}_{f,\mathcal{U}}(y^t)$. By applying Assumption 3 and the properties of λ, we thus have that

$$\|z - \xi(z_0, z)\| \le \sum_{j=1}^{N} \lambda_j \left\| \nabla_x m_j^t(y^t) - \nabla_x f(y^t, u^j) \right\| \le \kappa_g \Delta_t.$$

In the inner iteration t, our method employs a second-order model of the primal problem suggested by (2) around a current iterate y^t, but it replaces the uncertainty set $\mathcal{U}$ by the finite set $\{u^j : j \in J^{t,k}\} \subseteq \mathfrak{U}^k$. We then consider the extended trust-region subproblem

$$\min_{(z,d) \in \mathbb{R} \times \mathbb{R}^n} \left\{ z + \frac{1}{2} d^\top B^t d : \; F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} + (G^t)^\top d \le z\mathbf{e}, \; \|d\| \le \Delta_t \right\}, \tag{P}$$

where $\mathbf{e} \in \mathbb{R}^{|J^{t,k}|}$ denotes the vector of ones and $B^t \in \mathbb{S}_n$ is an approximating Hessian. We discuss in Section 6.1 a natural means for constructing B^t from interpolation conditions; here we remark that we do not require B^t to be positive definite for a given t but instead impose a uniform bound on $\|B^t\|$ as an algorithmic parameter κ_{mh}.

Our proposed algorithm applies a step acceptance test and a trust-region radius update typical of trust-region methods. That is, given a(n approximate) minimizer (z^t, d^t) of (P), and provided a certain condition (described below) on the generator set $\{u^j \in \mathfrak{U}^k : j \in J^{t,k}\}$ holds, we define the ratio of the actual decrease witnessed in $\Psi_{\mathfrak{U}^k}$ and the decrease predicted by the model in (P) by

$$\rho_t \triangleq \frac{\Psi_{\mathfrak{U}^k}(y^t) - \Psi_{\mathfrak{U}^k}(y^t + d^t)}{-(z^t + \frac{1}{2} d^{t\top} B^t d^t)}. \tag{10}$$

If ρ_t is sufficiently positive, then we accept the step d^t and possibly expand the trust-region radius; otherwise, the step d^t is rejected, and the trust-region radius is reduced.

We next discuss how to solve (P) approximately in order to achieve a sufficient reduction in the primal model. A dual measure χ is discussed in [11] for constrained trust-region subproblems; for our particular subproblem (P), this dual measure is readily shown to be

$$\chi_t \triangleq \min_{\lambda} \left\{ \|G^t \lambda\| + \lambda^\top (\Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} - F^t) : \lambda \geq 0,\ \mathbf{e}^\top \lambda = 1 \right\}. \tag{11}$$

In Section 6.1, we employ the relation that χ_t can be computed as the negative value of the Lagrangian dual to (P) when we set $B^t = \mathbf{0}_{n\times n}$ and $\Delta_t = 1$.

Before stating in Theorem 2 a sufficient decrease guarantee for (P), we show that χ_t is an upper bound on $-\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t)$, the negative approximate optimality measure, provided certain conditions hold. We first define the relevant subset of $\mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$ in (7) by

$$\mathcal{D}_{m^t,\mathfrak{U}^k_{J^{t,k}}}(y^t) \triangleq \mathbf{co} \left\{ \begin{bmatrix} \Psi_{\mathfrak{U}^k}(y^t) - m^t_j(y^t, u^j) \\ \nabla_x m^t_j(y^t) \end{bmatrix} : u^j \in \mathfrak{U}^k,\ j \in J^{t,k} \right\}. \tag{12}$$

Proposition 5 *For the convex subset $\mathcal{D}_{m^t,\mathfrak{U}^k_{J^{t,k}}}(y^t) \subseteq \mathcal{D}_{m^t,\mathfrak{U}^k}(y^t)$ in (12), provided there exists $(z_0, z) \in \mathcal{D}_{m^t,\mathfrak{U}^k_{J^{t,k}}}(y^t)$ satisfying $z_0 + \frac{1}{2}\|z\|^2 \leq 1$, then in the tth inner iteration of Algorithm 2,*

$$-\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t) \leq \chi_t.$$

Proof We have that

$$\begin{aligned} -\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t) &= \min_{z_0,z} \left\{ z_0 + \frac{1}{2}\|z\|^2 : (z_0, z) \in \mathcal{D}_{m^t,\mathfrak{U}^k}(y^t) \right\} \\ &\leq \min_{z_0,z} \left\{ z_0 + \frac{1}{2}\|z\|^2 : (z_0, z) \in \mathcal{D}_{m^t,\mathfrak{U}^k_{J^{t,k}}}(y^t) \right\} \\ &\leq \min_{z_0,z} \left\{ z_0 + \frac{1}{2}\|z\| : (z_0, z) \in \mathcal{D}_{m^t,\mathfrak{U}^k_{J^{t,k}}}(y^t) \right\} = \chi_t, \end{aligned}$$

where the last inequality comes from the assumption in the proposition statement, and the last equality is immediate from the definitions of $\mathcal{D}_{m^t,\mathfrak{U}^k_{J^{t,k}}}(y^t)$ and χ_t.

With this dual measure χ_t defined for our particular setting, we can now state the following theorem (adapted from [11]), which ensures that we can use Algorithm 3, provided in Appendix C, to obtain sufficient model reduction.

Theorem 2 *Suppose Algorithm 3 is called to solve an instance of (P) with algorithmic constants satisfying $0 < \kappa_{\rm ubs} < \kappa_{\rm lbs} < 1$, $\kappa_{\rm frd} \in (0, 1)$, and $\kappa_{\rm epp} \in (0, \frac{1}{2})$. Then:*

- *Algorithm 3 terminates in a finite number of iterations, and*
- *$[z^t; d^t]$ returned by Algorithm 3 satisfies the constraints in (P) and*

$$-\left(z^t + \frac{1}{2}\left(d^t\right)^\top B^t d^t \right) \geq \kappa_{\rm fcd}\chi_t \min\left\{ \frac{\chi_t}{\kappa_{\rm mh}}, \Delta_t, 1 \right\}, \tag{13}$$

where $\kappa_{\rm fcd} = \min\left\{\frac{1}{2}\kappa_{\rm ubs}\kappa_{\rm frd}, 2\kappa_{\rm ubs}(1 - \kappa_{\rm lbs})\right\}$ satisfies $\kappa_{\rm fcd} < 1$.

As a result of Theorem 2, we are justified in making the following assumption.

Assumption 4 *The constant $\kappa_{\rm fcd} \in (0, 1)$ is such that for any primal problem (P), we can find a (z^t, d^t) satisfying the relation (13) and constraints in (P).*

As for the selection of a set $J^{t,k}$ in inner iteration t and the mentioned criteria for step acceptance, we make use of a *manifold sampling loop* as in [19, 20]. We initialize $J^{t,k}$ with the active index (indices) $j^*(y^t)$ in the finite maximum function $\Psi_{\mathfrak{U}^k}(y^t)$; that is,

$$j^*(y^t) \in \operatorname*{argmax}_j \left\{ f(y^t, u^j) : u^j \in \mathfrak{U}^k \right\}.$$

We obtain a trial step (z^t, d^t) from the approximate solution of (P). Upon computing the value of $\Psi_{\mathfrak{U}^k}(y^t + d^t)$, we consider the active index (indices) $j^*(y^t + d^t)$ in $\Psi_{\mathfrak{U}^k}(y^t + d^t)$. If $j^*(y^t + d^t) \subseteq J^{t,k}$, we stop and perform the step acceptance test. Otherwise, we augment $J^{t,k}$ by $j^*(y^t + d^t)$ and compute a new trial step (z^t, d^t). Because there are finitely many elements in $\mathfrak{U}^k$, this augmentation, which we refer to as the *manifold sampling loop*, terminates in at most $|\mathfrak{U}^k|$ steps.

We provide a full statement of the proposed derivative-free algorithm for the solution of (1) as Algorithm 2, with algorithmic parameter assumptions appearing in Assumption 5.

Assumption 5 *The algorithmic parameters in Algorithm 2 satisfy the following:* $\kappa_{\mathrm{mh}} > 0$, $\gamma > 1$, $\eta_1 \in (0, 1)$, *and* $\eta_2 \in (0, 1/\kappa_{\mathrm{mh}})$.

A practical implementation of Algorithm 2 will be discussed in Section 6. A careful reader will note that performing the *acceptability test* in Line 23 is wasteful in terms of function evaluations; within the manifold sampling loop of inner iteration t, if a value of χ_t is encountered that will fail the acceptability test, then inner iteration t is guaranteed to be unacceptable, since χ_t is monotonically nonincreasing as $J^{t,k}$ is augmented. We have chosen to present the algorithm as is for clarity in our convergence analysis.

5 Convergence Analysis

The ultimate goal of our analysis is to prove the following theorem, which can be seen as a natural extension of Theorem 1.

Theorem 3 *Let Assumptions 1, 2, 3, 4, and 5 hold. Let* x^* *be an accumulation point of the sequence* $\{x^k\}_{k=0}^{\infty}$ *generated by Algorithm 2. Then* $\mathbf{0} \in \partial\Psi(x^*)$.

We proceed by analyzing Phase 1 of Algorithm 2, beginning in Line 9. Our strategy for proving Theorem 3 is to demonstrate that, for each outer iteration k, within a finite number of inner iterations t, we attain $-\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t) \le \chi_t < \epsilon_k$, where we have made use of Proposition 5. Then, by upper-bounding the inexact optimality measure $-\Theta_{\mathfrak{U}^k}(x^{k+1})$ used in Line 5 of Algorithm 1 in terms of the approximate optimality measure $-\tilde{\Theta}^t_{\mathfrak{U}^k}(x^{k+1})$, we will have proved the theorem, given the otherwise-identical algorithmic behavior of Algorithm 1 and Algorithm 2.

We begin by demonstrating that if $\chi_t \neq 0$ in Phase 1, then a successful iteration must occur within a finite number of inner iterations.

Lemma 2 *Let Assumptions 3, 4, and 5 hold. If Line 24 of Algorithm 2 is reached and*

$$\Delta_t < \min\left\{ \min\left\{ \frac{\kappa_{\mathrm{fcd}}(1-\eta_1)}{3\kappa_f + \frac{1}{2}\kappa_{\mathrm{mh}}}, \eta_2 \right\} \chi_t, 1 \right\},$$

then the test in Line 26 of Algorithm 2 is passed, and the inner iteration t *is successful.*

Algorithm 2: Derivative-Free Method of Outer Approximations

```
1  Choose $\{(\epsilon_k, \Omega^k)\}_{k=0}^{\infty}$ satisfying Assumption 2 and algorithmic parameters
   $(\kappa_{\mathrm{mh}}, \gamma, \eta_1, \eta_2)$ satisfying Assumption 5.
2  Choose starting point $x^0 \in \mathbb{R}^n$, $\mathfrak{U}^0 \subseteq \mathcal{U}$, and trust-region radius $\Delta_{\mathrm{init}} > 0$.
3  Set $k \leftarrow 0$.
4  while true do
5      $t \leftarrow 0$.
6      $y^t \leftarrow x^k$, $\Delta_t \leftarrow \Delta_{\mathrm{init}}$.
7      $\chi_t \leftarrow \infty$.
8      while $\chi_t > \epsilon_k$ do
9          (Phase 1) Choose set $J^{t,k}$ satisfying
           $\operatorname*{argmax}_{j=1,\ldots,|\mathfrak{U}^k|} f(y^t, u^j) \subseteq J^{t,k} \subseteq \{1, \ldots, |\mathfrak{U}^k|\}$.
10         while true do
11             (Manifold Sampling Loop) For each $j \in J^{t,k}$, construct a model $m_j^t$
               such that $m_j^t(s)$ is a fully linear model of $f(y, u^j)$ for all $y \in \mathcal{B}(y^t, \Delta_t)$.
12             Construct vector $F^t$ and matrix $G^t$ with entries and columns, respectively,
               corresponding to elements of $J^{t,k}$.
13             Choose an approximate Hessian $B^t$ satisfying $\|B^t\| \leq \kappa_{\mathrm{mh}}$.
14             Obtain $(z^t, d^t)$ satisfying the constraints in (P) and (13).
15             if $j^*(y^t + d^t) \subseteq J^{t,k}$ then
16                 Let $j^* \in j^*(y^t + d^t)$ be arbitrary.
17                 break (Go to Line 22).
18             else
19                 $J^{t,k} \leftarrow J^{t,k} \cup j^*(y^t + d^t)$.
20             end
21         end
22         $\rho_t \leftarrow 0$.
23         if $\Delta_t < \eta_2 \chi_t$ (acceptable iteration) then
24             Compute $\rho_t$ as in (10).
25         end
26         if $\rho_t > \eta_1$ (successful iteration) then
27             $y^{t+1} \leftarrow y^t + d^t$.
28             $\Delta_{t+1} \leftarrow \gamma \Delta_t$.
29         else
30             $\Delta_{t+1} \leftarrow \gamma^{-1} \Delta_t$.
31         end
32         $t \leftarrow t + 1$.
33     end
34     $x^{k+1} \leftarrow y^t$.
35     (Phase 2) Compute $u' \in \operatorname*{argmax}_{u \in \Omega^k} f(x^{k+1}, u)$.
36     Augment $\mathfrak{U}^{k+1} \leftarrow \mathfrak{U}^k \cup \{u'\}$.
37     $k \leftarrow k + 1$.
38 end
```

Proof Consider

$$
\begin{aligned}
|\rho_t - 1| &= \left| \frac{\Psi_{\mathfrak{U}^k}(y^t) - \Psi_{\mathfrak{U}^k}(y^t + d^t)}{-(z^t + \frac{1}{2} d^\top B^t d)} - 1 \right| \\
&= \left| \frac{\Psi_{\mathfrak{U}^k}(y^t) - \Psi_{\mathfrak{U}^k}(y^t + d^t) + z^t + \frac{1}{2} d^\top B^t d}{-(z^t + \frac{1}{2} d^\top B^t d)} \right| \triangleq \frac{\mathrm{num}_t}{\mathrm{denom}_t}.
\end{aligned}
\tag{14}
$$

By the constraints of (P),

$$z^t = \max_{j \in J^{t,k}} \left\{ F_j^t + (G_j^t)^\top d^t \right\} - \Psi_{\mathfrak{U}^k}(y^t), \tag{15}$$

where F_j^t denotes the jth entry of F^t and G_j^t denotes the jth column of G^t. Let $j^{**} \in J^{t,k}$ denote a maximizing index of the right-hand side in (15). Then, we can bound the numerator of (14) by

$$\begin{aligned} \mathrm{num}_t &= \left| F_{j^{**}}^t + G_{j^{**}}^{t\top} d^t - \Psi_{\mathfrak{U}^k}(y^t + d^t) + \tfrac{1}{2} d^\top B^t d \right| \\ &\le \left| F_{j^{**}}^t + G_{j^{**}}^{t\top} d^t - f(y^t + d^t, u^{j^{**}}) \right| \\ &\quad + \left| f(y^t + d^t, u^{j^{**}}) - \Psi_{\mathfrak{U}^k}(y^t + d^t) \right| + \left| \tfrac{1}{2} d^\top B^t d \right| \\ &\le \left(\kappa_f + \frac{1}{2} \kappa_{\mathrm{mh}} \right) \Delta_t^2 + \left| f(y^t + d^t, u^{j^{**}}) - \Psi_{\mathfrak{U}^k}(y^t + d^t) \right|, \end{aligned} \tag{16}$$

where in the last inequality we have used Assumption 3 and the bound on κ_{mh} from Assumption 5. Now, let j^* be shorthand for $j^*(y^t + d^t)$ as defined before (i.e., j^* is any maximizing index in $\Psi_{\mathfrak{U}^k}(y^t + d^t)$). Observe that, due to Assumption 3 and the definition of j^{**},

$$f(y^t + d^t, u^{j^{**}}) + \kappa_f \Delta_t^2 \ge F_{j^{**}}^t + G_{j^{**}}^{t\top} d^t \ge F_{j^*}^t + G_{j^*}^{t\top} d^t \ge f(y^t + d^t, u^{j^*}) - \kappa_f \Delta_t^2,$$

from which we conclude that

$$f(y^t + d^t, u^{j^*}) - f(y^t + d^t, u^{j^{**}}) \le 2\kappa_f \Delta_t^2. \tag{17}$$

Then, inserting (17) into (16), we obtain a bound on the numerator

$$\mathrm{num}_t \le \left(3\kappa_f + \frac{1}{2} \kappa_{\mathrm{mh}} \right) \Delta_t^2. \tag{18}$$

Combining (18) with the Cauchy decrease from Assumption 4, we can continue the bound in (14) by

$$|\rho_t - 1| = \left| \frac{\Psi_{\mathfrak{U}^k}(y^t) - \Psi_{\mathfrak{U}^k}(y^t + d^t) + z^t + \frac{1}{2} d^\top B^t d}{-(z^t + \frac{1}{2} d^\top B^t d)} \right| \le \frac{\left(3\kappa_f + \frac{1}{2} \kappa_{\mathrm{mh}} \right) \Delta_t^2}{\kappa_{\mathrm{fcd}} \chi_t \min\{\Delta_t, 1\}},$$

where we have used the fact that if Line 24 of Algorithm 2 is reached, then we must have (by using Assumption 5) $\Delta_t < \eta_2 \chi_t < \frac{\chi_t}{\kappa_{\mathrm{mh}}}$. Thus, by using the supposed bounds on Δ_t, we arrive at

$$|\rho_t - 1| \le \frac{\left(3\kappa_f + \frac{1}{2} \kappa_{\mathrm{mh}} \right) \Delta_t}{\kappa_{\mathrm{fcd}} \chi_t} < 1 - \eta_1;$$

and so, as desired, the test in Line 26 is passed.

We now show that for every pass through Phase 1, the sequence of trust-region radii tends to zero.

Lemma 3 *Let Assumptions 4 and 5 hold. For each pass through Phase 1 beginning in Line 8 in Algorithm 2, $\Delta_t \to 0$ as $t \to \infty$.*

Proof On successful (i.e., the test in Line 26 is passed) inner iterations t, we have that

$$\begin{aligned}\Psi_{\mathfrak{U}^k}(y^t) - \Psi_{\mathfrak{U}^k}(y^t + d^t) &> \eta_1[-(z^t + \tfrac{1}{2}d^\top B^t d)] \\ &\geq \eta_1 \kappa_{\mathrm{fcd}} \chi_t \min\left\{\frac{\chi_t}{\kappa_{\mathrm{mh}}}, \Delta_t, 1\right\} \\ &\geq \eta_1 \kappa_{\mathrm{fcd}} \chi_t \min\left\{\Delta_t, 1\right\} \\ &> \tfrac{\eta_1 \kappa_{\mathrm{fcd}}}{\eta_2} \min\left\{\Delta_t, \Delta_t^2\right\},\end{aligned}$$

where we have used the sufficient decrease from Assumption 4, the fact that $\eta_2 < 1/\kappa_{\mathrm{mh}}$ from Assumption 5, and the acceptability of successful iterations. If there are infinitely many successful iterations t, let them be indexed by $\{t_i\}_{i=0}^\infty$. Recall that, by definition, for any $u^{j'} \in \mathfrak{U}^k$, $\Psi_{\mathfrak{U}^k}(x^t) \geq f(x^t, u^{j'})$. Let $j^*(t)$ be shorthand for $j^*(y^t + d^t)$, an arbitrary active index in $\Psi_{\mathfrak{U}^k}(y^t + d^t)$. Since $\Psi_{\mathfrak{U}^k}(\cdot)$ is bounded below by Assumption 2, having infinitely many successful iterations implies that

$$\infty > \sum_{i=0}^{\infty} \Psi_{\mathfrak{U}^k}(y^{t_i}) - \Psi_{\mathfrak{U}^k}(y^{t_i} + d^{t_i}) \geq \tfrac{\eta_1 \kappa_{\mathrm{fcd}}}{\eta_2} \left(\sum_{i:\Delta_{t_i} \geq 1} \Delta_{t_i} + \sum_{i:\Delta_{t_i} < 1} \Delta_{t_i}^2 \right).$$

Note that there can only be finitely many iterations such that $\Delta_{t_i} \geq 1$, since otherwise the summation would be infinite, a contradiction. Thus, we can take

$$\kappa_4 = \frac{\eta_1 \kappa_{\mathrm{fcd}}}{\eta_2} \sum_{i:\Delta_{t_i} \geq 1} \Delta_{t_i} < \infty$$

and conclude that

$$\infty > \sum_{i=0}^{\infty} \Psi_{\mathfrak{U}^k}(y^{t_i}) - \Psi_{\mathfrak{U}^k}(y^{t_i} + d^{t_i}) \geq \kappa_4 + \frac{\eta_1 \kappa_{\mathrm{fcd}}}{\eta_2} \sum_{i:\Delta_{t_i} < 1} \Delta_{t_i}^2.$$

It follows that $\Delta_{t_i} \to 0$ for any infinite subsequence of successful iterations. Since Δ_{t_i} increases by a factor of γ on successful iterations, for any successful iterate t_i, $\gamma \Delta_{t_i} \geq \Delta_t \geq \Delta_{t_{i+1}}$ for all $t_i < t \leq t_{i+1}$. Thus, when the number of successful iterations is infinite, $\Delta_t \to 0$.

If there are only finitely many successful iterations, then there is a last successful iteration $t_{i'}$, and the algorithm monotonically decreases Δ_t for all iterations $t > t_{i'}$.

Thus, regardless of whether there are infinitely many or finitely many successful iterations, $\Delta_t \to 0$.

We now show that Phase 1 terminates in a finite number of iterations; that is, the loop beginning in Line 9 terminates finitely, returning an iterate x^{k+1} satisfying an approximate ϵ_k-stationarity condition.

Lemma 4 *Let Assumptions 2–5 hold. For the sequence $\{y^t, \Delta_t\}$ generated in the kth outer iteration of Algorithm 2, there exists $t(\epsilon_k)$ such that $-\tilde{\Theta}_{\mathfrak{U}^k}^{t(\epsilon_k)}(y^{t(\epsilon_k)}) \leq \chi_t < \epsilon_k$.*

Proof By Assumption 2.e, $\epsilon_k \le 1$ for all k, and so we conclude from Proposition 5 that it suffices to show the existence of $t(\epsilon_k)$ such that $\chi_{t(\epsilon_k)} < \epsilon_k$.

To arrive at a contradiction, suppose that $\chi_t \ge \epsilon_k$ for $t = 0, 1, \ldots$. By Lemma 2, any iteration satisfying $\Delta_t \le \min\{1, \kappa_3 \chi_t\}$ is successful, where

$$\kappa_3 = \min\left\{ \frac{\kappa_{\text{fcd}}(1-\eta_1)}{3\kappa_f + \frac{1}{2}\kappa_{\text{mh}}}, \eta_2 \right\}.$$

Under the contradiction hypothesis, this implies that every iteration t such that $\Delta_t \le \min\{1, \kappa_3\epsilon_k\}$ is successful, and so $\Delta_{t+1} = \gamma\Delta_t \ge \Delta_t$. Thus, $\Delta_t \ge \gamma^{-1}\min\{1, \kappa_3\epsilon_k\}$ for all t, contradicting Lemma 3.

We have now established in Lemma 4 that in each outer iteration k, Phase 1 terminates in a finite number of iterations $t(\epsilon_k)$. The following lemma describes the relationship between $\tilde{\Theta}^t_{\mathfrak{U}^k}(y^t)$ and $\Theta_{\mathfrak{U}^k}(y^t)$.

Lemma 5 *Let Assumptions 2, 3, and 5 hold, and suppose that in the tth inner iteration of Algorithm 2, $\chi_t \le \epsilon_k$ for some $\epsilon_k \in (0,1)$. Then,*

$$-\Theta_{\mathfrak{U}^k}(y^t) \le \epsilon_k + \kappa_g\eta_2\epsilon_k^2 + \frac{1}{2}\kappa_g^2\eta_2^2\epsilon_k^2.$$

Proof By supposition,

$$\min_{z_0,z}\left\{ z_0 + \frac{1}{2}\|z\|^2 : (z_0, z) \in \mathcal{D}_{m^t,\mathfrak{U}^k}(y^t) \right\} \le \epsilon_k;$$

let (z_0^*, z^*) denote a minimizer of this convex quadratic over a (convex) compact domain. By Lemma 1, there exists $(\xi_0(z_0^*, z^*), \xi(z_0^*, z^*)) \in \mathcal{D}_{f,\mathfrak{U}^k}(y^t)$ such that

$$\begin{aligned} z_0^* &= \xi_0(z_0^*, z^*) \\ \|z^* - \xi(z_0^*, z^*)\| &\le \kappa_g\Delta_t. \end{aligned}$$

By the reverse triangle inequality and then squaring both sides, this implies

$$\|\xi_0(z_0^*, z^*)\|^2 + \|\xi(z_0^*, z^*)\|^2 \le \|z^*\|^2 + 2\kappa_g\Delta_t\|z^*\| + \kappa_g^2\Delta_t^2. \tag{19}$$

Thus, we have

$$\begin{aligned} -\Theta_{\mathfrak{U}^k}(y^t) &= \min_{\xi_0,\xi}\left\{ \xi_0 + \frac{1}{2}\|\xi\|^2 : (\xi_0, \xi) \in \mathcal{D}_{f,\mathfrak{U}^k}(y^t) \right\} \\ &\le \xi_0(z_0^*, z^*) + \tfrac{1}{2}\|\xi(z_0^*, z^*)\|^2 \\ &\le z_0^* + \tfrac{1}{2}\|z^*\|^2 + \kappa_g\Delta_t\|z^*\| + \tfrac{1}{2}\kappa_g^2\Delta_t^2 && \text{by (5) and (19)} \\ &\le \epsilon_k + \kappa_g\Delta_t\|z^*\| + \tfrac{1}{2}\kappa_g^2\Delta_t^2 && \text{by definition of } (z_0^*, z^*) \\ &\le \epsilon_k + \kappa_g\eta_2\chi_t^2 + \tfrac{1}{2}\kappa_g^2\eta_2^2\chi_t^2 && \text{since } \Delta_k < \eta_2\|z^*\|,\ \|z^*\| \le \chi_t \\ &\le \epsilon_k + \kappa_g\eta_2\epsilon_k^2 + \tfrac{1}{2}\kappa_g^2\eta_2^2\epsilon_k^2 && \text{by supposition,} \end{aligned}$$

which is what we intended to show.

We immediately obtain the desired proof of Theorem 3:

Proof **(of Theorem 3)** By Lemma 5, the iterate x^{k+1} returned at the end of Phase 1 of Algorithm 2 satisfies

$$-\Theta_{\mathfrak{U}^k}(x^{k+1}) \le \epsilon_k' \triangleq \epsilon_k + \kappa_g\eta_2\epsilon_k^2 + \frac{1}{2}\kappa_g^2\eta_2^2\epsilon_k^2.$$

Since $\{\epsilon_k'\} \to 0$ by virtue that $\{\epsilon_k\} \to 0$ due to Assumption 2.e, the result holds due to Theorem 1, where we simply replace $\{\epsilon_k\}$ with $\{\epsilon_k'\}$.

6 Practical Considerations

We present here a number of considerations intended to facilitate practical implementation of Algorithm 2.

6.1 Strong Duality of Trust-Region Subproblem

We first complete a previous claim about the computability of our dual measure χ_t in (11).

Consider the Lagrangian function of (P),

$$\mathcal{L}(z,d,\lambda,\mu) \triangleq z + \frac{1}{2}d^\top B^t d + \lambda^\top \left(F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} + (G^t)^\top d - z\mathbf{e}\right) + \frac{\mu}{2}(d^\top d - \Delta_t^2).$$

The dual function

$$g(\lambda,\mu) \triangleq \min_{z,d} \mathcal{L}(z,d,\lambda,\mu) \tag{20}$$

is unbounded below unless $\lambda^\top \mathbf{e} = 1$ and $B^t + \mu I_n \succeq 0$. Provided that (λ,μ) satisfies these dual constraints, (20) is equivalent to

$$g(\lambda,\mu) = \lambda^\top (F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e}) - \tfrac{\mu}{2}\Delta_t^2 + \min_d \left\{ \frac{1}{2} d^\top (B^t + \mu I_n) d + (G^t\lambda)^\top d \right\}. \tag{21}$$

From (21), we see that $g(\lambda,\mu)$ is unbounded below unless there exists d such that $(B^t + \mu I_n)d = -G^t\lambda$. One can easily show that when $B^t + \mu I_n \succeq \mathbf{0}$, then every vector in the set $D^* \triangleq \{d : (B^t + \mu I_n)d = -G^t\lambda\}$ is a global minimizer in (21); thus, even if D^* is a nontrivial subspace of $\mathbb{R}^n$, it follows that $g(\lambda,\mu)$ is well defined.

Collecting these observations, we arrive at the Lagrangian dual of (P):

$$\begin{array}{rr} \displaystyle\max_{(\mu,\lambda,v)\in\mathbb{R}\times\mathbb{R}^{|J^{t,k}|}\times\mathbb{R}^n} & \lambda^\top (F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e}) + \tfrac{1}{2} v^\top G^t \lambda - \mu \frac{\Delta_t^2}{2} \\ \text{subject to} & \lambda \geq \mathbf{0} \\ & \mu \geq 0 \\ & \mathbf{e}^\top \lambda = 1 \\ & B^t + \mu I_n \succeq \mathbf{0} \\ & (B^t + \mu I_n) v = -G^t \lambda. \end{array} \tag{D}$$

We remark that, in light of our comment on $g(\lambda,\mu)$ being well defined even when D^* is nonsingleton, maximizing over $v \in \mathbb{R}^n$ in (D) is equivalent to $\max\limits_{\lambda\geq\mathbf{0},\mu\geq 0} g(\lambda,\mu)$.

We make the following important observation.

Proposition 6 *The optimization problems* (P) *and* (D) *are strongly dual; that is, their optimal values are attained and are equal.*

Proof Let $val(P)$ denote the optimal value of (P), and let $val(D)$ denote the optimal value of (D). Note that (P) is always feasible, as is evident from the feasibility of $(z,d) = (0,\mathbf{0})$. Moreover, a minimum is attained by Weierstrass's theorem. By weak duality, we always have $val(D) \leq val(P)$, so it suffices to show $val(P) \leq val(D)$.

Observe that the Mangasarian-Fromowitz constraint qualification holds for any problem (P). Indeed, as an extreme case, suppose every constraint of (P) is active at a minimum (z^*, d^*); all other cases follow from similar analysis. The gradient of the trust-region constraint at (z^*, d^*) is $[0,\, d^*]^\top$, and the gradient of the jth linear inequality is $[-1,\, g^j]^\top$. Then, letting $\epsilon > 0$ be arbitrary, the vector $v = \left[\max_{j \in J^{t,k}} (-g^j)^\top d^* + \epsilon,\, -d^*\right]^\top$ satisfies $[0,\, d^*]^\top v < 0$ and $[-1,\, g^j]^\top v < 0$ for all $j \in J^{t,k}$. Thus, the Karush-Kuhn-Tucker conditions hold at (z^*, d^*); and so, in particular, complementarity conditions hold at (z^*, d^*). The complementarity conditions for (P) imply that

$$\lambda^{*\top}\left(F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} + G^{t\top} d^* - z^*\mathbf{e}\right) = 0 \tag{22}$$

and

$$\mu^*\left(\frac{d^{*\top} d^*}{2} - \frac{\Delta_t^2}{2}\right) = 0. \tag{23}$$

So,

$$\begin{aligned}
val(P) &= z^* + \tfrac{1}{2} d^{*\top} B^t d^* \\
&= z^* + \tfrac{1}{2} d^{*\top}(B^t + \mu^* I) d^* - \tfrac{\mu}{2} d^{*\top} d^* \\
&= z^* - \tfrac{1}{2} d^{*\top} G^t \lambda^* - \tfrac{\mu^*}{2} d^{*\top} d^* \\
&= \lambda^{*\top}(F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} + G^{t\top} d^*) - \tfrac{1}{2} d^{*\top} G^t \lambda^* - \tfrac{\mu^*}{2} d^{*\top} d^* \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{by (22) and } \mathbf{e}^\top \lambda^* = 1 \\
&= \lambda^{*\top}(F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e}) + \tfrac{1}{2} d^{*\top} G^t \lambda^* - \tfrac{\mu^*}{2} d^{*\top} d^* \\
&\le \lambda^{*\top}(F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e}) + \max_v \left\{ \frac{1}{2} v^\top G^t \lambda^* : (B^t + \mu^* I) v = -G^t \lambda^* \right\} - \frac{\mu^*}{2} d^{*\top} d^* \\
&= \lambda^{*\top}(F^t - \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e}) + \max_v \left\{ \frac{1}{2} v^\top G^t \lambda^* : (B^t + \mu^* I) v = -G^t \lambda^* \right\} - \frac{\mu^*}{2} \Delta_t^2 \\
&= val(D),
\end{aligned}$$

where the second-to-last equality is because if $\mu^* = 0$, then the equality is trivially true, while if $\mu^* > 0$, then $d^{*\top} d^* = \Delta_t^2$ by (23).

By setting $B^t = \mathbf{0}_{n \times n}$ and $\Delta_t = 1$ in (D), we obtain $val(D) = -\chi_t$; this follows by noting that for any λ^* in this setting, the corresponding optimal value of μ is $\mu^* = \|G^t \lambda^*\|$. Thus, computing χ_t can be seen as a special case of solving (P).

Having established this equivalence, we believe that the following, listed in the order in which they are reached in Algorithm 2, are three important considerations in a practical implementation of Algorithm 2.

1. Selection of $\{\Omega^k\}_{k=0}^\infty$
2. Computation of F^t and G^t (i.e., construction of the models m_j^t)
3. Selection of B^t

We address each of these items in the remainder of this section.

6.2 Selecting $\{\Omega^k\}_{k=0}^\infty$

Although convenient for proving convergence, preselecting the sequence $\{\Omega^k\}_{k=0}^\infty$ at the start of Algorithm 2 is not desirable. In practice, Ω^k instead should be chosen adaptively, for example, based on function values $\{f(x^{k+1},u) : u \in \mathfrak{U}^k\}$ already obtained in Phase 1.

As a baseline, and in the interest of satisfying Assumption 2, we propose that selection of Ω^k include a variation of the following stochastic sampling scheme.

Let $\phi : \mathcal{U} \to \mathbb{R}_{++}$ be a probability density function that is strictly positive on its support $\mathcal{U}$. Then, when Phase 2 is reached in the kth iteration of Algorithm 2, we select a sample size $s_k \in \mathbb{N}$ so that s_k is $\mathcal{O}(k)$, and we set $\Omega^k = \{\omega^1, \dots, \omega^{s_k}\}$, for samples ω^i drawn according to ϕ for each $i = 1, \dots, s_k$. We can easily show that, because ϕ is strictly positive on $\mathcal{U}$, Assumption 2.d is satisfied with probability one. In Section 7, we experiment with combinations of stochastic and deterministic selection schemes for adaptively selecting Ω^k.

6.3 Constructing m_j^t

We propose that incorporating second-order information be passed to a selection of B^t and hence that m_j^t simply be a linear model of $f(y^t, u^j)$ that is formed by interpolating on a set of sufficiently affinely independent points $P \subset \mathcal{B}(y^t, \Delta_t)$ such that $|P| = n+1$ and $y^t \in P$. This is enough to satisfy Assumption 3; see, for instance, [12, Theorems 2.11 and 2.12]. In practice we reuse previously evaluated points as much as possible. If $n+1$ affinely independent points contained in $\mathcal{B}(y^t, \Delta_t)$ are not available in our algorithm's history for some $u^j \in \mathfrak{U}^k$, then we perform evaluations of $f(p^i, u^j)$ for at most n additional points p^i until we are again assured that m_j^t is a fully linear model of $f(y^t + s, u^j)$ for all $s \in \mathcal{B}(y^t, \Delta_t)$ (see, e.g., [24, Figure 4.2]).

Consequently, upon exiting Line 11 of Algorithm 2, we have a sufficiently affinely independent set $P \subset \mathcal{B}(y^t, \Delta_t)$ such that $|P| = n+1$ and $y^t \in P$, with the additional property that we have already obtained function evaluations $\{f(p^i, u^j)\}$ for all $p^i \in P$ and for all $j \in J^{t,k}$. This will be valuable immediately in Section 6.4.

6.4 Choosing B^t

We propose to construct B^t in such a way that the model objective in (P) interpolates (or regresses) the set of sample points $P = \{p^1, p^2, \dots, p^{|P|}\}$ used in the construction of $\{m_j^t : j \in J^{t,k}\}$ in Line 11 of Algorithm 2. That is, given the current iterate y^t, letting $p^i = y^t + s^i$ for $i = 1, \dots, |P|$, and letting Φ_Q denote the quadratic polynomial basis defined by

$$\Phi_Q(v) \triangleq \left[\frac{1}{2}v_1^2, \dots, \frac{1}{2}v_n^2, v_1v_2, \dots, v_2v_3, \dots, v_{n-1}v_n\right],$$

we obtain coefficients α_Q by solving (in the least-squares sense)

$$\begin{bmatrix} \Phi_Q(p^1) \\ \vdots \\ \Phi_Q(p^{|P|}) \end{bmatrix} \alpha_Q = \begin{bmatrix} \Psi_{\mathfrak{U}(J^{t,k})}(p^1) - \max\limits_{j=1,\dots,|J^{t,k}|} \left\{ F_j^t + (G_j^t)^\top s^1 \right\} \\ \vdots \\ \Psi_{\mathfrak{U}(J^{t,k})}(p^{|P|}) - \max\limits_{j=1,\dots,|J^{t,k}|} \left\{ F_j^t + (G_j^t)^\top s^{|P|} \right\} \end{bmatrix}, \qquad (24)$$

where F_j^t denotes the jth entry of F^t, G_j^t denotes the jth column of G^t, and $\mathfrak{U}(J^{t,k}) = \{u^j \in \mathfrak{U}^k : j \in J^{t,k}\}$.

Such a Hessian fitting technique requires no additional function evaluations, since all the necessary evaluations to compute $\Psi_{\mathfrak{U}(J^{t,k})}(p)$ for all $p \in P$ were performed in Line 11 of Algorithm 2. Replacing each instance of $\Psi_{\mathfrak{U}(J^{t,k})}$ in (24) with $\Psi_{\mathfrak{U}^k}$ could involve many additional function evaluations, and we have found that Hessian fitting via (24) performs well in practice.

7 Numerical Results

Here we present our results from testing a GNU Octave/Matlab implementation of a variant of Algorithm 2.

We begin by outlining differences between our implementation and the theoretical algorithm. We use algorithmic parameters $\gamma = 2, \eta_1 = .001, \kappa_{\text{mh}} = 1000$. In a departure from the theory, we set $\eta_2 = \infty$ (i.e., we effectively do not perform an acceptability test) because we found that performing an acceptability test hurts efficiency as measured by the number of function evaluations. For all problems we will consider, we set the default initial trust region as $\Delta_{\text{init}} = 1$ and define the tolerance sequence by $\epsilon_k = 2^{-k}$.

In this implementation, we elect to solve (P) (and hence, by our remarks in Section 6, the subproblem for the calculation of χ_t) via Matlab's fmincon. We do not perform a feasibility test.

When constructing B^t, we do as proposed in Section 6, with the following caveat: If $\|B^t\|_F > \kappa_{\text{mh}}$, where $\|\cdot\|_F$ denotes the Frobenius norm of a matrix, then we simply set $B^t = \mathbf{0}_{n\times n}$, effectively yielding a problem without curvature. In future work we will seek a reasonable subproblem to replace that in (24), but which appropriately constrains $\|B^t\|_F$ so that this naive removal of curvature need not occur.

7.1 Test Functions

We employ two classes of functions in order to illustrate features of Algorithm 2. These functions are of low dimension, in keeping with the typical situation that for high-dimensional problems, significantly more evaluations can be required by derivative-free algorithms than are required by their derivative-based counterparts.

The first function we consider is a two-dimensional polynomial:

$$\begin{aligned} g(x) = {} & 2x_1^6 - 12.2x_1^5 + 21.2x_1^4 - 6.4x_1^3 - 4.7x_1^2 + 6.2x_1 + x_2^6 - 11x_2^5 + 43.3x_2^4 \\ & -74.8x_2^3 + 56.9x_2^2 - 10x_2 - 0.1x_1^2 + x_2^2 + 0.4x_1^2x_2 + 0.4x_2^2x_1 - 4.1x_1x_2. \end{aligned} \tag{25}$$

A now-standard (to our knowledge, first appearing in [7]) robust optimization problem is obtained by considering g in (25) in the presence of *implementation errors*; that is, we consider the problem

$$\min_{x\in\mathbb{R}^2} \Psi_{\mathcal{U}_\alpha}(x) \triangleq \min_{x\in\mathbb{R}^2} \max_{u:\|u\|_2\leq\alpha} f(x,u) \triangleq \min_{x\in\mathbb{R}^2} \max_{u:\|u\|_2\leq\alpha} g(x+u),$$

where $\alpha \geq 0$ is a parameter, with $\alpha = 0.5$ being the value considered by a number of robust optimization studies (e.g., [6, 7, 13]); see Figure 1. We remark that

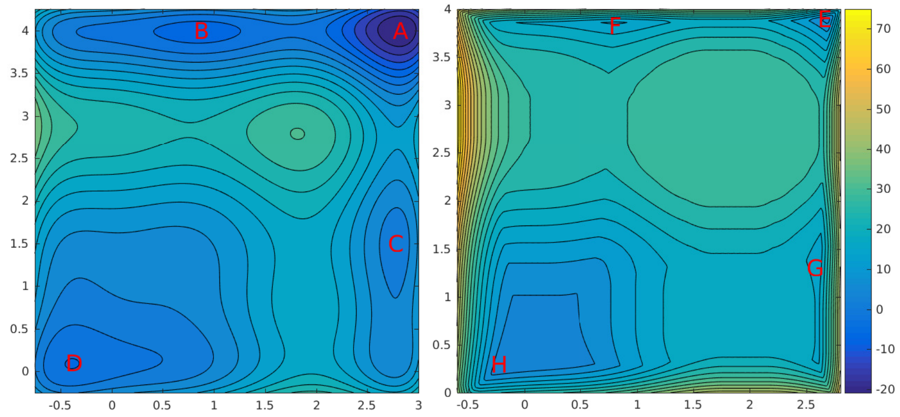

Fig. 1 For the function (25), contour plots of (left) the nominal function $g(x) = f(x, 0)$ and (right) $\Psi_{\mathcal{U}_\alpha}(x)$ for $\alpha = 0.5$. Local minima are labeled by the letters A,B,C,D and E,F,G,H for the nominal and robust problems, respectively. We remark that A and H are global minimizers of the nominal and robust problems, respectively.

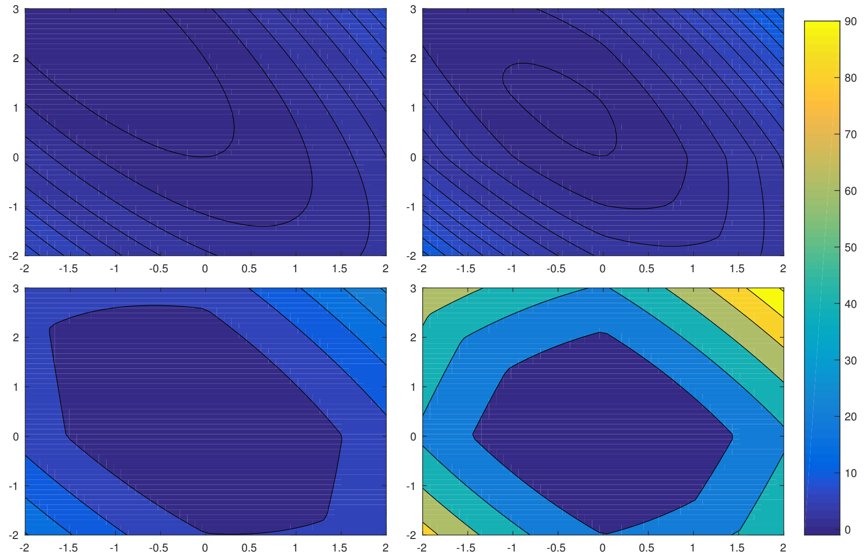

Fig. 2 For fixed $\hat{L}, \hat{b}$, contours of $\Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}(x)$ in (26) for $\alpha = 0, 0.125, 0.5, 2$ (top left, top right, bottom left, bottom right, respectively).

implementation errors are a special type of uncertainty; in particular, if we have previously evaluated a point $y \in \mathbb{R}^2$ (i.e., we have computed the value $g(y)$), then for any $x \in \mathbb{R}^2$, we automatically obtain $f(x, y - x) = g(x + (y - x)) = g(y)$. An efficient algorithm would exploit this fact; since this paper addresses general robust problems of the form (1), we deliberately treat $f : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ as a black-box function.

The second family of functions we consider are the biquadratics defined in the following way. Let $\hat{L} \in \mathbb{R}^{n\times n}$ denote a lower-triangular matrix with nonzero diagonal entries, and let $\hat{b} \in \mathbb{R}^n$. With this data $(\hat{L}, \hat{b})$ we define the minimax problem

$$\min_{x\in\mathbb{R}^n} \Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}(x) \triangleq \min_{x\in\mathbb{R}^n} \max_{(L,b)\in\mathcal{U}_\alpha(\hat{L},\hat{b})} \frac{1}{2}x^\top L^\top L x + b^\top x, \tag{26}$$

where, for $\alpha > 0$, the uncertainty set $\mathcal{U}_\alpha(\hat{L}, \hat{b})$ is defined by

$$\left\{(L,b) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n : \, |L_{ij} - \hat{L}_{ij}| \le \alpha, \, \forall i \ge j; \; L_{ij} = 0, \, \forall i < j; \; |b_i - \hat{b}_i| \le \alpha, \forall i\right\}.$$

Problems defined by (26) have several appealing properties. First, by using $x = \mathbf{0}$ in (26), it follows that $\Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}(x) \le 0$ for all $\alpha \ge 0$. Moreover, because the nominal problem defined by $(L, b) = (\hat{L}, \hat{b})$ is a strictly convex quadratic, $\Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}(x)$ is bounded below. It follows that the family of functions satisfies Assumption 1. Furthermore, (26) is a useful benchmark for Line 6 of Algorithm 1; for fixed x, the solution u' to the subproblem in Line 6 with $\Omega^k = \mathcal{U}_\alpha(\hat{L}, \hat{b})$ can be obtained, for instance, from a solution to the bound-constrained convex maximization problem

$$\max_{(\ell,b)\in\mathbb{R}^{n^2}\times\mathbb{R}^n} \left\{\frac{1}{2}\ell^\top \left(I_n \otimes xx^\top\right)\ell + x^\top b \, : \, (\mathbf{mat}(\ell), b) \in \mathcal{U}_\alpha(\hat{L}, \hat{b})\right\}, \tag{27}$$

where $\mathbf{mat}(\ell)$ denotes the matrix L obtained from "unrolling" the vector ℓ and $\otimes$ is the Kronecker product of matrices. The particular structure of (27) allows for a globally optimal solution to be obtained efficiently, despite the fact that (27) is a convex maximization problem; we provide details on the solution of (27) in Appendix D. Moreover, problems of the form (26) are useful for benchmarking because the function $\Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}$ is convex (a maximum of convex functions). Since the stationarity measure $\Theta_{\mathcal{U}_\alpha(\hat{L},\hat{b})}$ is intractable to compute and nontrivial to approximate, here we use values of $\Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}$ to measure progress.

Figure 2 shows a two-dimensional example of (26), where we have fixed a random $(\hat{L}, \hat{b})$ and varied the parameter α.

7.2 Illustrative Results

We first illustrate in Figure 3 desirable properties of our implementation.

We initialize our implementation from the four local minima of the nominal function (25) shown in Figure 1. We initialize $\mathfrak{U}^0$ as $\{\pm 0.5e_i : i = 1, 2\}$, where e_i denotes the ith column of the identity matrix I_2. We note that with this initialization, the iterates of our implementation manage to identify the basin of the global minimizer ("H") of $\Psi_{\mathcal{U}_\alpha}(x)$. In all four cases, in the generation of Figure 3, a budget of 250 evaluations of the function f was imposed.

We now consider the biquadratics of the form (26) in a "low-dimensional" setting, in which the dimension of x is $n = 2$ (and so $\mathcal{U}_\alpha \subset \mathbb{R}^5$), and in a "high-dimensional" setting, in which $n = 8$ (and so $\mathcal{U}_\alpha \subset \mathbb{R}^{44}$). We randomly generate nominal parameters $\hat{L}$ and $\hat{b}$ in the following way. Each entry of $\hat{L}$ is drawn from a uniform distribution on $[-1, 1]$; by inspection, we ensure that all generated values

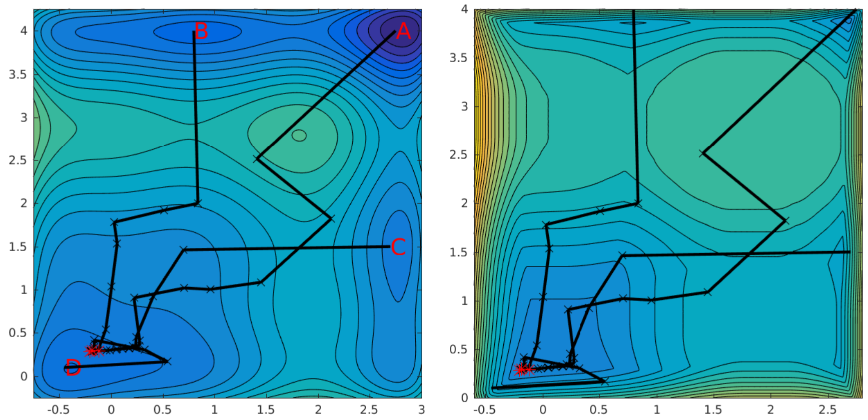

Fig. 3 Trajectories of our implementation initialized at local minimizers of the nominal function (25) (compare with Figure 1).

are nonzero. We then randomly generate $\hat{b}$ from the subspace spanned by the eigenvectors of $\hat{L}^\top \hat{L}$ corresponding to eigenvalues greater than 0.01. This procedure ensures that the system $\hat{L}^\top \hat{L}x = \hat{b}$ is sufficiently well-conditioned; thus, we can set x^0 as the optimal solution to the nominal problem without encountering excessively large initial values of $\Psi_{\mathcal{U}_\alpha(\hat{L},\hat{b})}(x^0)$. In these experiments, we set $\mathfrak{U}^0$ as the nominal values of the parameters, $(\hat{L}, \hat{b})$.

In each of our two settings ($n = 2, n = 8$), we choose the parameter $\alpha = \frac{1}{n}$ and then generate 30 random instances and run variants of our implementation. As described below, these variants differ only in the way that Phase 2 is performed and are meant to test the recommendations made in Section 6.2.

- **(Gaussian RBF)** If there are fewer than $m+1$ points in $\mathfrak{U}^k$, then we perform a rank-revealing QR decomposition to complete $\mathfrak{U}^k$ into an affinely independent set $\bar{\mathfrak{U}}^k$. We obtain a model $m^f(u)$ by interpolating a Gaussian radial basis function (RBF) with a linear tail on the set of points $\{(x^{k+1}, u) : u \in \bar{\mathfrak{U}}^k\}$; see, for example, [25]. We remark that if $|\mathfrak{U}^k| \geq m + 1$, then the function values $\{f(x^{k+1}, u) : u \in \mathfrak{U}^k\}$ will already have been performed in Phase 1, and so no function evaluations are performed in this step. If $|\mathfrak{U}^k| < m + 1$, then we perform an additional $m + 1 - |\mathfrak{U}^k|$ function evaluations. We then set Ω^k to be an approximate solution (as found by fmincon, initialized at $\operatorname{argmax}_{u\in\bar{\mathfrak{U}}^k} f(x^{k+1}, u)$) to the problem $\max\limits_{u\in\mathcal{U}_\alpha} m^f(u)$.
- **(Uniform Random Sampling)** We set Ω^k to be a set of $\lceil \beta m \rceil$ random samples uniformly generated from $\mathcal{U}_\alpha$, where $\beta > 0$. After some tuning, we determined $\beta = 2$ was a good value for these problems.
- **(Optimal Phase 2)** Although clearly not an option in general black-box optimization, we set Ω^k to be the solution to (27). This is strictly for benchmarking purposes.

For a performance metric, we can compute $\Psi(x)$ in postprocessing exactly at every successful trial step visited by our implementation by solving (27). In

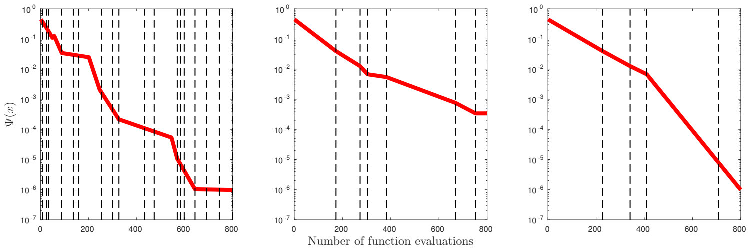

Fig. 4 Trajectories of $\Psi(x)$ as a function of the number of function evaluations for solving (26) with the same $(\hat{L}, \hat{b})$ as shown in the bottom left of Figure 2. We compare Gaussian RBF (left), uniform random sampling with $\beta = 2$ (middle), and Optimal Phase 2 (right). The vertical dashed lines indicate the end of a pass through Phase 1.

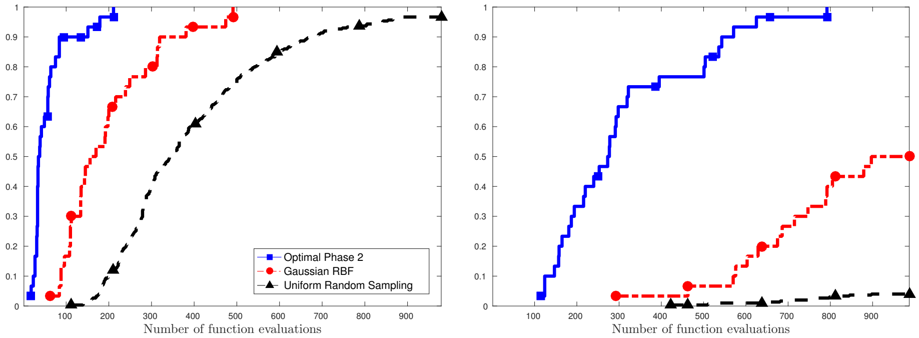

Fig. 5 Data profiles for $\Psi(x)$ for low-dimensional ($n = 2, m = 5, \alpha = 0.5$) biquadratics (26). Levels of accuracy shown are $\tau = 10^{-1}$ and $\tau = 10^{-5}$ in the left and right plots, respectively.

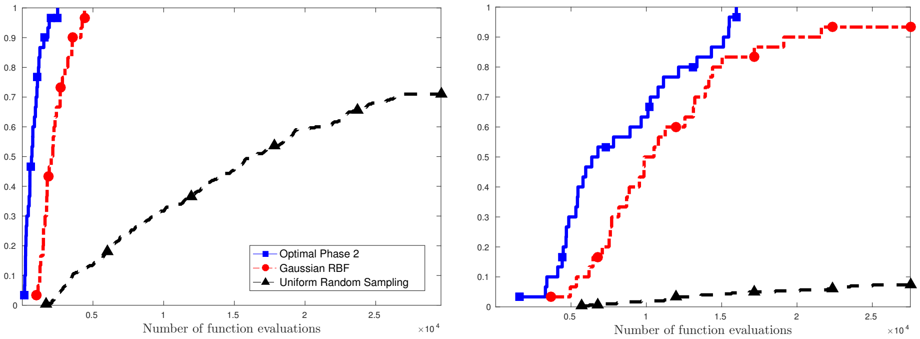

Fig. 6 Data profiles for $\Psi(x)$ for low-dimensional ($n = 8, m = 44, \alpha = 0.125$) biquadratics (26). Levels of accuracy shown are $\tau = 10^{-1}$ and $\tau = 10^{-5}$ in the left and right plots, respectively.

Figure 4, we demonstrate that our choice of budget is appropriate by showing true values of $\Psi(x)$ produced for the $\alpha = 0.5$ instance shown in Figure 2 as a function of the number of function evaluations.

In Figures 5 and 6, we show the results of these experiments by using data profiles [21]. Data profiles show the empirical distribution function, in terms of the number of evaluations of f, of the problems solved by each solver. Following [21], we consider a problem to be solved to a level $\tau \geq 0$ provided that a x is evaluated satisfying

$$\Psi_{\mathcal{U}}(x^0) - \Psi_{\mathcal{U}}(x) \geq (1-\tau)\left(\Psi_{\mathcal{U}}(x^0) - \Psi_{\mathcal{U}}(x^{\text{best}})\right),$$

where x^0 is a starting point common to all of the solvers and $\Psi_{\mathcal{U}}(x^{\text{best}})$ denotes the best $\Psi_{\mathcal{U}}$ value obtained by any of the solvers within the computational budget provided.

In the case of uniform random sampling, we performed 30 trials; we show here the empirical cumulative distribution across all these trials and problems.

We observe in Figures 5 and 6 that while some efficiency is lost by not being able to perform Phase 2 optimally, the Gaussian RBF interpolation strategy is reasonable in a derivative-free setting, particularly in the lower accuracy tests represented by $\tau = 10^{-1}$.

Acknowledgments

This material was based upon work supported by the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, applied mathematics program under Contract No. DE-AC02-06CH11357.

A Optimality Measure Properties

Here we collect proofs for Section 2.

A.1 Proof of Proposition 2

For fixed $\hat{x} \in \mathbb{R}^n$, we have that $\theta(\hat{x}, h)$ from (3) can be written as

$$\begin{aligned}\theta(\hat{x}, h) &= \max_{(\xi_0,\xi)\in\mathcal{E}(\hat{x})} (-\xi_0 + \Psi(\hat{x})) + \langle \xi, h\rangle + \frac{1}{2}\|h\|^2 - \Psi(\hat{x}) \\ &= \max_{(\xi_0,\xi)\in\mathcal{E}(\hat{x})} -\xi_0 + \langle \xi, h\rangle + \frac{1}{2}\|h\|^2,\end{aligned}$$

which is a maximization of a linear function over

$$\mathcal{E}(\hat{x}) \triangleq \cup_{u\in\mathcal{U}} \begin{bmatrix} \Psi(\hat{x}) - f(\hat{x}, u) \\ \nabla_x f(\hat{x}, u) \end{bmatrix} \subseteq \mathbf{co}\mathcal{E}(\hat{x}) = \mathcal{D}_{f,\mathcal{U}}(\hat{x}) \subseteq \mathbb{R}^{n+1}.$$

Thus, its optimal value is equal to the optimal value of

$$\max_{(\xi_0,\xi)\in\mathcal{D}_{f,\mathcal{U}}(\hat{x})} -\xi_0 + \langle \xi, \hat{h}\rangle + \frac{1}{2}\|\hat{h}\|^2 \tag{28}$$

since an extreme point of $\mathcal{D}_{f,\mathcal{U}}(\hat{x})$, which is necessarily in $\mathcal{E}(\hat{x})$ by definition of the convex hull, is an optimal solution of (28). Thus, we have established that

$$\Theta(\hat{x}) = \min_{h\in\mathbb{R}^n} \max_{(\xi_0,\xi)\in\mathcal{D}_{f,\mathcal{U}}(\hat{x})} -\xi_0 + \langle \xi, h\rangle + \frac{1}{2}\|h\|^2. \tag{29}$$

Letting $b(h, (\xi_0, \xi)) \triangleq -\xi_0 + \langle \xi, h\rangle + \frac{1}{2}\|h\|^2$, the function involved in the minimax expression of (29), we note that

- $b(h, (\xi_0, \xi))$ is continuous on $\mathbb{R}^n \times \mathbb{R}^{n+1}$;
- $b(h, (\hat{\xi}_0, \hat{\xi}))$ is strictly convex in h for any $(\hat{\xi}_0, \hat{\xi}) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$;
- $b(\hat{h}, (\xi_0, \xi))$ is concave in (ξ_0, ξ) for any $\hat{h} \in \mathbb{R}^n$;
- $\mathcal{D}_{f,\mathcal{U}}(\hat{x})$ is, by definition, a convex set; and
- $b(h, (\xi_0, \xi)) \to \infty$ as $\|h\| \to \infty$ uniformly in $(\xi_0, \xi) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$.

Thus, the conditions of von Neumann's theorem apply, and so we conclude that (29) is equivalent to

$$\Theta(\hat{x}) = \max_{(\xi_0,\xi)\in\mathcal{D}_{f,\mathcal{U}}(\hat{x})} \min_{h\in\mathbb{R}^n} -\xi_0 + \langle \xi, h \rangle + \frac{1}{2}\|h\|^2. \tag{30}$$

Now, for a fixed $(\hat{\xi}_0, \hat{\xi}) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$, the solution to the unconstrained convex inner minimization problem of (30) satisfies (by sufficient and necessary first-order conditions) $h = -\hat{\xi}$. Thus, the inner minimization in (30) can be replaced with $-\hat{\xi}_0 - \frac{\|\hat{\xi}\|^2}{2}$, yielding the desired result

$$\Theta(\hat{x}) = \max_{(\xi_0,\xi)\in\mathcal{D}_{f,\mathcal{U}}(\hat{x})} -\xi_0 - \frac{1}{2}\|\xi\|^2.$$

A.2 Proof of Proposition 3

Clearly, $\xi_0 = \Psi(\hat{x}) - f(\hat{x}, u) \geq 0$ for all $(\xi_0, \xi) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$. Combined with the nonnegativity of norms, it follows immediately from the definition of $\Theta(\hat{x})$ in (5) that $\Theta(\hat{x}) = 0$ if and only if $\mathbf{0} \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$. Thus, it suffices to show that $\mathbf{0} \in \partial\Psi(\hat{x})$ if and only if $\mathbf{0} \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$.

Suppose that $\mathbf{0} \in \partial\Psi(\hat{x})$. Let $u^*(\hat{x}) \in \mathcal{U}^*(\hat{x})$, where we have defined

$$\mathcal{U}^*(\hat{x}) \triangleq \operatorname*{argmax}_{u\in\mathcal{U}} f(\hat{x}, u).$$

Then, for any such $u^*(\hat{x})$, $\Psi(\hat{x}) - f(\hat{x}, u^*(\hat{x})) = 0$. It follows that the set

$$D^*(\hat{x}) \triangleq \{(\xi_0, \xi) : \xi_0 = 0, \xi \in \partial\Psi(\hat{x})\}$$

satisfies $D^*(\hat{x}) \subseteq \mathcal{E}(\hat{x}) \subseteq \mathcal{D}_{f,\mathcal{U}}(\hat{x})$. Thus, $\mathbf{0} \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$.

Now suppose that $\mathbf{0} \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$. By Caratheodory's theorem and the convex hull definition of $\mathcal{D}_{f,\mathcal{U}}(\hat{x})$ in (4), there exist $q \leq n+2$; $u^1, \ldots, u^q \in \mathcal{U}$; and $\{\lambda \in \mathbb{R}^q_+ : \lambda_1 + \cdots + \lambda_q = 1\}$ such that

$$\mathbf{0} = \sum_{j=1}^{q} \lambda_j \begin{bmatrix} \Psi(\hat{x}) - f(\hat{x}, u^j) \\ \nabla_x f(\hat{x}, u^j) \end{bmatrix}. \tag{31}$$

Clearly, $\Psi(\hat{x}) - f(\hat{x}, \hat{u}) = \operatorname*{argmax}_{u\in\mathcal{U}} f(\hat{x}, u) - f(\hat{x}, \hat{u}) \geq 0$ for all $\hat{u} \in \mathcal{U}$. Thus, projecting the convex combination (31) into its first coordinate, we must have that all q vectors satisfy $\Psi(\hat{x}) - f(\hat{x}, u^j) = 0$; that is,

$$u^j \in \operatorname*{argmax}_{u\in\mathcal{U}} f(\hat{x}, u) \quad \text{for} \quad j = 1, \ldots, q. \tag{32}$$

Likewise, projecting (31) into its last n coordinates,

$$\mathbf{0} = \sum_{j=1}^{q} \lambda_j \nabla_x f(\hat{x}, u^j). \tag{33}$$

Together, (32) and (33) imply that $\mathbf{0} \in \partial\Psi(\hat{x})$.

A.3 Proof of Proposition 4

For completeness, we state the definition of a continuous set-valued mapping.

Definition 1 Consider a sequence of sets $\{S_j\}_{j=0}^{\infty} \subset \mathbb{R}^n$.

1. The point $x^* \in \mathbb{R}^n$ is a *limit point of* $\{S_j\}$ provided $dist(x^*, S_j) \to 0$.
2. The point $x^* \in \mathbb{R}^n$ is a *cluster point of* $\{S_j\}$ if there exists a subsequence $\mathcal{K}$ such that $dist(x^*, S_j) \to_{\mathcal{K}} 0$.
3. We denote the set of limit points of $\{S_j\}$ by $\liminf S_j$ and refer to it as the *inner limit.*
4. We denote the set of cluster points of $\{S_j\}$ by $\limsup S_j$ and refer to it as the *outer limit.*

Definition 2 We say that a set-valued mapping $\Gamma : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is

1. *outer semicontinuous (o.s.c.) at* $\hat{x}$ provided for all sequences $\{x^j\} \to \hat{x}$, $\limsup \Gamma(x^j) \subseteq \Gamma(\hat{x})$,
2. *inner semicontinuous (i.s.c.) at* $\hat{x}$ provided for all sequences $\{x^j\} \to \hat{x}$, $\liminf \Gamma(x^j) \supseteq \Gamma(\hat{x})$, and
3. *continuous at* $\hat{x}$ provided Γ is o.s.c. and i.s.c. at $\hat{x}$.

Without proof, we state Corollary 5.3.9 from [22].

Proposition 7 *Suppose that $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p$ is continuous and that $\Gamma : \mathbb{R}^n \to 2^{\mathbb{R}^m}$ is a continuous set-valued mapping. Then, the set-valued mapping $G : \mathbb{R}^n \to 2^{\mathbb{R}^p}$ defined by*

$$G(x) \triangleq \mathbf{co}\,\{g(x,u) : u \in \Gamma(x)\} \tag{34}$$

is continuous.

By using Proposition 7, we get the following intermediate result needed to prove continuity of Θ.

Proposition 8 *Let Assumption 1 hold; then, the set-valued mapping $\mathcal{D}_{f,\mathcal{U}}(\cdot) : \mathbb{R}^n \to 2^{\mathbb{R}^{n+1}}$ is continuous.*

Proof We look to (34) in Proposition 7 as a template. In the definition of $\mathcal{D}_{f,\mathcal{U}}(\cdot)$, $\Gamma(x) = \mathcal{U}$ for all $x \in \mathbb{R}^n$, and as such, $\mathcal{U}$ is trivially a continuous set-valued mapping. We have only to show that $D : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^{n+1}$ defined by

$$D(x,u) \triangleq \begin{bmatrix} \Psi(x) - f(x,u) \\ \nabla_x f(x,u) \end{bmatrix}$$

is continuous. Continuity follows since, by Assumption 1, $\Psi(x) - f(x,u)$ is a continuous function on $\mathbb{R}^n \times \mathcal{U}$, and $\nabla_x f(x,u) : \mathbb{R}^n \times \mathcal{U} \to \mathbb{R}^n$ is a Lipschitz continuous function on $\mathbb{R}^n \times \mathcal{U}$.

We can now prove Proposition 4:

Proof Consider the equivalent form of Θ from Proposition 2 in (5),

$$\Theta(\hat{x}) = \max_{(\xi_0,\xi) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})} q(\xi_0, \xi),$$

where we have defined the concave quadratic $q(\xi_0,\xi) \triangleq -\xi_0 - \dfrac{1}{2}\|\xi\|^2$.

Let $\hat{x}$ be arbitrary, and let $\{x^j\}_{j=0}^{\infty}$ be an arbitrary sequence satisfying $x^j \to \hat{x}$. For $j = 0, 1, \ldots$, let (ξ_0^j, ξ^j) be any $(\xi_0^j, \xi^j) \in \mathcal{D}_{f,\mathcal{U}}(x^j)$ such that $\Theta(x^j) = -\xi_0^j - \dfrac{1}{2}\|\xi^j\|^2$.

The sequence $\{x^j\}_{j=0}^{\infty}$ is bounded (it is convergent by assumption); we can also show that despite the arbitrary selection, there exists $M \geq 0$ such that $\|(\xi_0^j, \xi^j)\| \leq M$ uniformly for $j = 0, 1, \ldots$. To see this, suppose instead that $\|(\xi_0^j, \xi^j)\| \to \infty$. Then, since $\mathcal{D}_{f,\mathcal{U}}(\hat{x})$ is a compact set, there exists $M \geq 0$ such that $\max\limits_{(\xi_0,\xi) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})} \|(\xi_0, \xi)\| = M$. By our contradiction hypothesis, there exists $\underline{j}$ sufficiently large so that $\|(\xi_0^j, \xi^j)\| > 2M$ for all $j \geq \underline{j}$. From Proposition 8, $\mathcal{D}_{f,\mathcal{U}}(\cdot)$ is a continuous set-valued mapping. Thus, for any $\epsilon > 0$, there exists $\underline{j}(\epsilon) \geq \underline{j}$ sufficiently large so that $\textbf{dist}\left((\xi_0^j, \xi^j), \mathcal{D}_{f,\mathcal{U}}(\hat{x})\right) < \epsilon$ for all $j > \underline{j}(\epsilon)$; this means that $\|(\xi_0^j, \xi^j)\| \leq M + \epsilon$. This is impossible for all $\epsilon \in [0, M]$, yielding a contradiction.

Thus, since $\|(\xi_0^j, \xi^j)\| \leq M$ for $j = 0, 1, \ldots$ and because $q(\cdot)$ is a continuous function of (ξ_0, ξ), $\limsup_{j\to\infty} q(\xi_0^j, \xi^j)$ exists by the Bolzano-Weierstrass theorem. Let $\mathcal{K}$ denote a subsequence witnessing

$$\lim_{j\in\mathcal{K}} q(\xi_0^j, \xi^j) = \limsup_{j\to\infty} q(\xi_0^j, \xi^j),$$

and let $(\hat{\xi}_0, \hat{\xi}) = \lim_{j\in\mathcal{K}} (\xi_0^j, \xi^j)$ denote the corresponding accumulation point. Again using the fact that $\mathcal{D}_{f,\mathcal{U}}(\cdot)$ is o.s.c., we conclude that $(\hat{\xi}_0, \hat{\xi}) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$. Using the definition of $\Theta(\hat{x})$, we have

$$\Theta(\hat{x}) \geq q(\hat{\xi}_0, \hat{\xi}) = \lim_{j\in\mathcal{K}} q(\xi_0^j, \xi^j) = \limsup_{j\to\infty} q(\xi_0^j, \xi^j) = \limsup_{j\to\infty} \Theta(x^j). \tag{35}$$

As written, (35) means that $\Theta(\cdot)$ is upper semicontinuous. We now demonstrate that $\Theta(\cdot)$ is also lower semicontinuous, which will complete the proof of the continuity of $\Theta(\cdot)$. To establish a contradiction, we suppose that there exist $\hat{x} \in \mathbb{R}^n$ and a sequence $\{x^j\}_{j=0}^\infty$ satisfying $x^j \to \hat{x}$ such that $\Theta(x^j)$ exists for all j and

$$\lim_{j\to\infty} \Theta(x^j) < \Theta(\hat{x}). \tag{36}$$

Let $(\hat{\xi}_0, \hat{\xi}) \in \mathcal{D}_{f,\mathcal{U}}(\hat{x})$ satisfy $\Theta(\hat{x}) = q(\hat{\xi}_0, \hat{\xi})$. Since $\mathcal{D}_{f,\mathcal{U}}(\hat{x})$ is a continuous set-valued mapping by Proposition 8, there exists a $(\xi_0^j, \xi^j) \in \mathcal{D}_{f,\mathcal{U}}(x^j)$ satisfying $\Theta(x^j) = q(\xi_0^j, \xi^j)$ for all $j = 0, 1, \ldots$ such that $(\xi_0^j, \xi^j) \to (\hat{\xi}_0, \hat{\xi})$. Since $q(\cdot)$ is a continuous function in (ξ_0, ξ), $\lim_{j\to\infty} q(\xi_0^j, \xi^j) = q(\hat{\xi}_0, \hat{\xi})$. Thus, by using the contradiction hypothesis (36), we have

$$q(\hat{\xi}_0, \hat{\xi}) = \lim_{j\to\infty} q(\xi_0^j, \xi^j) = \lim_{j\to\infty} \Theta(x^j) < \Theta(\hat{x}) = q(\hat{\xi}_0, \hat{\xi}),$$

the desired contradiction.

B Convergence of Inexact Method of Outer Approximation

We now establish intermediate results needed to prove Theorem 1.

For brevity of notation, we use the following shorthand for the quadratic objective that appears in the definition of the optimality measure (6):

$$q_{\hat{\mathcal{U}}}(x, h) \triangleq \max_u \left\{ f(x, u) + \langle \nabla_x f(x, u), h \rangle + \frac{1}{2}\|h\|^2 \,:\, u \in \hat{\mathcal{U}} \right\}. \tag{37}$$

Consistent with our previous notation, we write $q(x, h)$ in (37) in the case where $\hat{\mathcal{U}} = \mathcal{U}$.

Lemma 6 *Let $\mathcal{S} \subset \mathbb{R}^n$ be a bounded subset. Suppose Assumptions 1 and 2 hold, and let $L \in [0, \infty)$ be a Lipschitz constant valid for $f(\cdot, \cdot)$ and $\nabla_x f(\cdot, \cdot)$ on $\mathcal{S} \times \mathcal{U}$. Then, there exists $\kappa_1 < \infty$ such that for all $x \in \mathcal{S}$ and for all $k = 0, 1, \ldots,$*

$$|\Psi_{\Omega^k}(x) - \Psi(x)| \leq \kappa_1 \delta(k).$$

Moreover, for the $\delta : \mathbb{N} \to \mathbb{R}$ from Assumption 2, there exists $\kappa_2 \in (\kappa_1, \infty)$ such that

$$|\Theta_{\Omega^k}(x) - \Theta(x)| \leq \kappa_2 \delta(k).$$

Proof Since $\Omega^k \subseteq \mathcal{U}$, we have that $\Psi_{\Omega^k}(\hat{x}) \leq \Psi(\hat{x})$ for all $\hat{x} \in \mathcal{S}$ and all $k = 0, 1, \ldots$.

Fix $\hat{x} \in \mathcal{S}$ and $u^*(\hat{x}) \in \mathcal{U}^*(\hat{x})$. Then, by definition of Ψ, $\Psi(\hat{x}) = f(\hat{x}, u^*(\hat{x}))$. By Assumption 2, for all k, there exists $[u^*(\hat{x})]' \in \Omega^k$ and $\kappa_0 > 0$ such that $\|u^*(\hat{x}) - [u^*(\hat{x})]'\| \leq \kappa_0 \delta(k)$. Thus,

$$\Psi_{\Omega^k}(\hat{x}) \geq f(\hat{x}, [u^*(\hat{x})]') \geq f(\hat{x}, u^*(\hat{x})) - L\kappa_0\delta(k) = \Psi(\hat{x}) - L\kappa_0\delta(k), \tag{38}$$

proving the first part of the lemma, with $\kappa_1 = L\kappa_0$.

For the second part, let $\hat{x} \in \mathcal{S}$ and $\hat{h} \in \mathbb{R}^n$ be arbitrary. By the definition of q in (37),

$$\min_{h \in \mathbb{R}^n} q_{\Omega^k}(\hat{x}, h) \leq q_{\Omega^k}(\hat{x}, \hat{h}) \leq q(\hat{x}, \hat{h})$$

for any $k = 0, 1, \ldots$. Since $\hat{h}$ was arbitrary, we can replace it with a minimizer of the convex $q(\hat{x}, \cdot)$; that is,

$$\min_{h \in \mathbb{R}^n} q_{\Omega^k}(\hat{x}, h) \leq \min_{h' \in \mathbb{R}^n} q(\hat{x}, h'). \tag{39}$$

Observing that Θ in (2) and Θ_{Ω^k} in (6) can be written, respectively, as

$$\begin{aligned} \Theta(\hat{x}) &= \min_{h \in \mathbb{R}^n} q(\hat{x}, h) - \Psi(\hat{x}) \\ \Theta_{\Omega^k}(\hat{x}) &= \min_{h \in \mathbb{R}^n} q_{\Omega^k}(\hat{x}, h) - \Psi_{\Omega^k}(\hat{x}), \end{aligned}$$

we conclude from (38) and (39) that

$$\begin{aligned} \Theta_{\Omega^k}(\hat{x}) &= \min_{h \in \mathbb{R}^n} q_{\Omega^k}(\hat{x}, h) - \Psi_{\Omega^k}(\hat{x}) \\ &\leq \min_{h \in \mathbb{R}^n} q(\hat{x}, h) - \Psi_{\Omega^k}(\hat{x}) \\ &= \Theta(\hat{x}) + \Psi(\hat{x}) - \Psi_{\Omega^k}(\hat{x}) \\ &\leq \Theta(\hat{x}) + L\kappa_0\delta(k). \end{aligned} \tag{40}$$

Denote the minimizer of the $\Theta_{\Omega^k}(\hat{x})$ by

$$h_k(\hat{x}) \triangleq \operatorname*{argmin}_{h \in \mathbb{R}^n} q_{\Omega^k}(\hat{x}, h) - \Psi_{\Omega^k}(\hat{x}).$$

Then, from the dual characterization of $\Theta_{\Omega^k}(\hat{x})$ in Proposition 2, we have that

$$h_k(\hat{x}) \in \left\{ -\xi : (\xi_0, \xi) \in \mathcal{D}_{f, \Omega^k}(\hat{x}) \right\} = \left\{ -\nabla f(\hat{x}, u) : u \in \Omega^k \right\}. \tag{41}$$

By Assumption 1 and since we supposed $\mathcal{S}$ and Ω^k are bounded, $\nabla_x f(\cdot, u)$ is continuous over $\mathcal{S}$ for each $u \in \Omega^k$; furthermore, by (41), there exists $M \in [0, \infty)$ such that $\|h_k(x)\| \leq M$ for all $x \in \mathcal{S}$. Let $u^*(\hat{x}) \in \mathcal{U}$ be a maximizer in the definition of $q(\hat{x}, h_k(\hat{x}))$ in (37) such that

$$q(\hat{x}, h_k(\hat{x})) = f(\hat{x}, u^*(\hat{x})) + \langle \nabla_x f(\hat{x}, u^*(\hat{x})), h_k(\hat{x}) \rangle + \frac{1}{2}\|h_k(\hat{x})\|^2. \tag{42}$$

By Assumption 2, for all k, there exists $[u^*(\hat{x})]' \in \Omega^k$ such that $\|u^*(\hat{x}) - [u^*(\hat{x})]'\| \leq \kappa_0\delta(k)$. Combining that with the Lipschitz continuity of Assumption 1, we obtain both

$$\left| f(\hat{x}, u^*(\hat{x})) - f\left(\hat{x}, [u^*(\hat{x})]'\right) \right| \leq L\kappa_0\delta(k)$$

and

$$\begin{aligned} |\langle \nabla_x f(\hat{x}, u^*(\hat{x})) - \nabla_x f(\hat{x}, [u^*(\hat{x})]'), h_k(\hat{x}) \rangle| &\leq \|\nabla_x f(\hat{x}, u^*(\hat{x})) - \nabla_x f(\hat{x}, [u^*(\hat{x})]')\| \, \|h_k(\hat{x})\| \\ &\leq MLc\delta(k). \end{aligned}$$

Combining these Lipschitz bounds with (42), we obtain

$$\begin{aligned} q_{\Omega^k}(\hat{x}, h_k(\hat{x})) &\geq f(\hat{x}, [u^*(\hat{x})]') + \langle \nabla_x f(\hat{x}, [u^*(\hat{x})]'), h_k(\hat{x}) \rangle + \frac{1}{2}\|h_k(\hat{x})\|^2 \\ &\geq q(\hat{x}, h_k(\hat{x})) - (M+1)L\kappa_0\delta(k). \end{aligned} \tag{43}$$

Using the definition of $\Theta_{\Omega^k}(\hat{x})$, we can rewrite (43) as

$$\Theta_{\Omega^k}(\hat{x}) + \Psi_{\Omega^k}(\hat{x}) \geq q(\hat{x}, h_k(\hat{x})) - (M+1)L\kappa_0\delta(k). \tag{44}$$

Likewise, by using the fact that $\Theta(\hat{x}) = \operatorname*{argmin}_{h\in\mathbb{R}^n} q(\hat{x}, h) - \Psi(\hat{x}) \leq q(\hat{x}, h_k(\hat{x})) - \Psi(\hat{x})$, (44) is equivalent to

$$\Theta_{\Omega^k}(\hat{x}) \geq \Theta(\hat{x}) + \Psi(\hat{x}) - \Psi_{\Omega^k}(\hat{x}) - (M+1)L\kappa_0\delta(k). \tag{45}$$

Inserting the bound from (38) into (45), we obtain

$$\Theta_{\Omega^k}(\hat{x}) \geq \Theta(\hat{x}) - (M+2)L\kappa_0\delta(k). \tag{46}$$

Combining the bounds in (40) and (46), we have proved the second part of the lemma, with $\kappa_2 = (M+2)L\kappa_0$, since $\kappa_2 > \kappa_1 = L\kappa_0$.

The next lemma demonstrates that, under our assumptions, the accumulation points x^* of a sequence $\{x^k\}$ generated by Algorithm 1 satisfy (on the same subsequence K defining the accumulation) $\Psi_{\mathfrak{U}^k}(x^k) \to_K \Psi(x^*)$.

Lemma 7 *Suppose that Assumptions 1 and 2 hold and that both*

1. $\{x^k\}_{k=0}^{\infty} \subset \mathbb{R}^n$ *and*
2. $\mathfrak{U}^k \subseteq \Omega^k$ *are constructed recursively with* $\mathfrak{U}^0 \neq \emptyset$, $\mathfrak{U}^0 \subseteq \mathcal{U}$, *and* $\mathfrak{U}^{k+1} = \mathfrak{U}^k \cup \{u'\}$, *where* $u' \in (\Omega^{k+1})^*(x^{k+1})$.

If x^* *is an accumulation point of* $\{x^k\}_{k=0}^{\infty}$ *(i.e., for some infinite subset* $\mathcal{K} \subset \mathbb{N}$, $x^k \to_{\mathcal{K}} x^*$*), then* $\Psi_{\mathfrak{U}^k}(x^k) \to_{\mathcal{K}} \Psi(x^*)$.

Proof For any $k \in \{1, 2, \dots\}$, let $\underline{k} \triangleq \max\{k' \in \mathcal{K} : k' \leq k\}$. Then, by our recursive construction, for any k, $u^{\underline{k}} \in \mathfrak{U}^k$. Since $\mathfrak{U}^k \subseteq \mathcal{U}$ for $k = 0, 1, \dots,$

$$\Psi(x^k) \geq \Psi_{\mathfrak{U}^k}(x^k) \geq f(x^k, u^{\underline{k}}). \tag{47}$$

By the triangle inequality,

$$|\Psi_{\Omega^{\underline{k}}}(x^{\underline{k}}) - \Psi(x^*)| \leq |\Psi_{\Omega^{\underline{k}}}(x^{\underline{k}}) - \Psi(x^{\underline{k}})| + |\Psi(x^{\underline{k}}) - \Psi(x^*)|. \tag{48}$$

Because $x^k \to_{\mathcal{K}} x^*$ and because $\Psi(\cdot)$ is a continuous function as a result of Assumption 1, the second summand in (48) satisfies $|\Psi(x^{\underline{k}}) - \Psi(x^*)| \to 0$. By Lemma 6 and the continuity of $\Psi(\cdot)$, we also conclude that the first summand in (48) satisfies $|\Psi_{\Omega^{\underline{k}}}(x^{\underline{k}}) - \Psi(x^{\underline{k}})| \to 0$. Thus,

$$\Psi_{\Omega^{\underline{k}}}(x^{\underline{k}}) \to \Psi(x^*). \tag{49}$$

Since from Assumption 1, $f(\hat{x}, u)$ is a uniformly continuous function in u over a compact set, and since $\|x^k - x^{\underline{k}}\| \to 0$ (by accumulation), we have

$$|f(x^k, u^{\underline{k}}) - f(x^{\underline{k}}, u^{\underline{k}})| \to 0.$$

By definition, $\Psi_{\Omega^{\underline{k}}}(x^{\underline{k}}) = f(x^{\underline{k}}, u^{\underline{k}})$, and so the above can be written

$$|f(x^k, u^{\underline{k}}) - \Psi_{\Omega^{\underline{k}}}(x^{\underline{k}})| \to 0. \tag{50}$$

It follows immediately from (49) and (50) that $f(x^k, u^{\underline{k}}) \to \Psi(x^*)$. So, by (47) and an application of the sandwich theorem, we conclude $\Psi_{\mathfrak{U}^k}(x^k) \to_{\mathcal{K}} \Psi(x^*)$, as we intended to show.

By using Lemma 7, we can now give a proof of Theorem 1.

Proof **(of Theorem 1)** Recalling the definition of $q_{\widehat{\mathcal{U}}}$ in (37), and since $\mathfrak{U}^k \subseteq \mathcal{U}$ for $k = 0, 1, \ldots,$ we have that for all k and for all $\hat{h} \in \mathbb{R}^n$, $q_{\mathfrak{U}^k}(x^k, \hat{h}) \leq q(x^k, \hat{h})$. Then, recalling the definition of $\Theta_{\mathfrak{U}^k}$ in (6), we have that

$$\begin{aligned} \Theta_{\mathfrak{U}^k}(x^k) &= \min_{h \in \mathbb{R}^n} q_{\mathfrak{U}^k}(x^k, h) - \Psi_{\mathfrak{U}^k}(x^k) \\ &\leq \min_{h \in \mathbb{R}^n} q(x^k, h) - \Psi_{\mathfrak{U}^k}(x^k) \\ &= \Theta(x^k) + \Psi(x^k) - \Psi_{\mathfrak{U}^k}(x^k). \end{aligned} \tag{51}$$

By using the criteria imposed on $\Theta_{\mathfrak{U}^k}(x^{k+1})$ in Line 5 of Algorithm 1 and (51), we have that for $k = 0, 1, \ldots,$

$$-\epsilon_k \leq \Theta_{\mathfrak{U}^k}(x^{k+1}) \leq \Theta(x^{k+1}) + \Psi(x^{k+1}) - \Psi_{\mathfrak{U}^{k+1}}(x^{k+1}). \tag{52}$$

Let $\mathcal{K}$ be a subsequence defining the accumulation $\{x^k\} \to_{\mathcal{K}} x^*$. Taking the limit with respect to $\mathcal{K}$ in (52), we obtain

$$\begin{array}{llr} & \lim\limits_{k \in \mathcal{K}} -\epsilon_k \leq \lim\limits_{k \in \mathcal{K}} \left[\Theta(x^{k+1}) + \Psi(x^{k+1}) - \Psi_{\mathfrak{U}^{k+1}}(x^{k+1})\right] & \\ \iff & 0 \leq \lim\limits_{k \in \mathcal{K}} \Theta(x^{k+1}) + \lim\limits_{k \in \mathcal{K}} \Psi(x^{k+1}) - \lim\limits_{k \in \mathcal{K}} \Psi_{\mathfrak{U}^{k+1}}(x^{k+1}) & \text{by } \epsilon_k \to 0 \\ \iff & 0 \leq \lim\limits_{k \in \mathcal{K}} \Theta(x^{k+1}) + \Psi(x^*) - \lim\limits_{k \in \mathcal{K}} \Psi_{\mathfrak{U}^{k+1}}(x^{k+1}) & \text{by continuity of } \Psi \\ \iff & 0 \leq \lim\limits_{k \in \mathcal{K}} \Theta(x^{k+1}) + \Psi(x^*) - \Psi(x^*) & \text{by Lemma 7} \\ \iff & 0 \leq \lim\limits_{k \in \mathcal{K}} \Theta(x^{k+1}) \leq 0 & \text{by Proposition 1.} \end{array}$$

By the continuity of Θ from Proposition 4, the result follows from the sandwich theorem.

C Availability of a Generalized Cauchy Point

We refer the reader to [11, Chapter 12.2] for a detailed discussion of generalized Cauchy decrease in trust-region subproblems with convex (here, linear) constraints, but we provide some necessary details here, beginning with the following definition.

Definition 3 Let $p(r) : \mathbb{R} \to \mathbb{R}^{n+1}$ denote the projection $\mathcal{P}_{\mathcal{C}}([-r; \mathbf{0}])$, where

$$\mathcal{C} = \left\{[z; d] : G^{t\top} d - z\mathbf{e} \leq \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} - F^t\right\}.$$

Use the notation $p(r) = [p_z(r); p_d(r)]$ to indicate the separation of $p(r)$ into the scalar z component and the n-dimensional d component. Then, the *generalized Cauchy point for* (P) is defined as $p(r^*)$, where

$$r^* = \operatorname*{argmin}_r \left\{p_z(r) + \frac{1}{2} p_d(r)^\top B^t p_d(r) : 0 \leq r \leq \Delta_t\right\}.$$

The generalized Cauchy point is the global minimizer of the objective in (P) restricted to an arc described by the projected steepest descent direction at $(z, d) = (0, \mathbf{0})$. Algorithm 3 (see [11, Algorithm 12.2.2]) computes an *approximate* generalized Cauchy point for (P) via a Goldstein-type line search. The notation $\mathcal{T}_{\mathcal{C}}(y)$ denotes the tangent cone to a convex set $\mathcal{C}$ at a point y (and we remark that, given a linear polytope $\mathcal{C}$, this set is easily computable).

We further remark that the computation of $p(r)$ for a given r involves the solution of the convex quadratic program

$$\min_{s_z, s_d} \left\{(r + s_z)^2 + \|s_d\|^2 : G^{t\top} s_d - s_z\mathbf{e} \leq \Psi_{\mathfrak{U}^k}(y^t)\mathbf{e} - F^t\right\}.$$

Although we anticipate that Algorithm 3 has benefits in many real-world settings, here it is merely of theoretical convenience, and we do not use it in the implementation tested.

Algorithm 3: Goldstein-type line search for generalized Cauchy point of (P)

```
Choose constants $0 < \kappa_{\text{ubs}} < \kappa_{\text{lbs}} < 1$, $\kappa_{\text{frd}} \in (0, 1)$, $\kappa_{\text{epp}} \in (0, \frac{1}{2})$.
Set $r_{\min} \leftarrow 0, r_{\max} \leftarrow \infty, r_0 = \Delta_t, j \leftarrow 0$.
while true do
    Compute $p(r_j) = [p_z(r_j); p_d(r_j)]$ as in Definition 3.
    if $\|p_d(r_j)\| > \Delta_t$ or $p_z(r_j) + \frac{1}{2} p_d(r_j)^\top B^t p_d(r_j) > \kappa_{\text{ubs}} p_z(r_j)$ then
        $r_{\max} \leftarrow r_j$.
    else if $\|p_d(r_j)\| < \kappa_{\text{frd}} \Delta_t$ and $p_z(r_j) + \frac{1}{2} p_d(r_j)^\top B^t p_d(r_j) < \kappa_{\text{lbs}} p_z(r_j)$ and
    $\|\mathcal{P}_{\mathcal{T}_C(p(r_j))}([-1; \mathbf{0}])\| > \frac{\kappa_{\text{epp}} |p_z(r_j)|}{\Delta_t}$ then
        $r_{\min} \leftarrow r_j$.
    else
        $[z^t; d^t] \leftarrow p(r_j)$. break
    if $r_{\max} = \infty$ then
        $r_{j+1} \leftarrow 2r_j$.
    else
        $r_{j+1} \leftarrow \frac{1}{2}(r_{\min} + r_{\max})$
    $j \leftarrow j + 1$.
    return $[z^t; d^t]$
end
```

D Global Maximization of (27)

First, we remark that the objective of (27) is separable with respect to the variables L and b. Thus, it is evident that the optimal value of b is given by

$$b_i^* = \begin{cases} \hat{b}_i - \alpha, & \text{if } x_i < 0 \\ \hat{b}_i + \alpha, & \text{otherwise} \end{cases} \qquad i = 1, \ldots, n.$$

We now consider the optimal value of L. After deleting rows and columns of $I_n \otimes xx^\top$ corresponding to the entries L_{ij} where $L_{ij} = 0$, we are left with a matrix of the form

$$\begin{bmatrix} x_{\bar{1}} x_{\bar{1}}^\top & \mathbf{0} & \cdots & & \mathbf{0} \\ \mathbf{0} & x_{\bar{2}} x_{\bar{2}}^\top & \mathbf{0} & \cdots & \mathbf{0} \\ \vdots & \mathbf{0} & \ddots & & \vdots \\ & \vdots & & & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & x_{\bar{n}} x_{\bar{n}}^\top \end{bmatrix},$$

where $x_{\bar{i}}$ denotes the truncated vector $[x_1, \ldots, x_i]$. Exploiting this block structure, the maximization of the quadratic decomposes into n bound-constrained quadratic maximization problems of the form

$$\max_{\ell \in \mathbb{R}^i} \left\{ \frac{1}{2} \ell^\top \left(x_{\bar{i}} x_{\bar{i}}^\top \right) \ell : \, |\ell_j - \hat{\ell}_j| \leq \alpha, \; j = 1, \ldots, i \right\} \tag{53}$$

for $i = 1, \ldots, n$. In turn, solving (53) is equivalent to solving the problem

$$\max_{\ell \in \mathbb{R}^i} \left\{ |x_{\bar{i}}^\top \ell| : \, |\ell_j - \hat{\ell}_j| \leq \alpha, \; j = 1, \ldots, i \right\}, \tag{54}$$

which can be cast as a mixed-integer linear program with exactly one binary variable; that is, solving (54) to global optimality entails the solution of two linear programs with $\mathcal{O}(i)$ variables and $\mathcal{O}(i)$ constraints each. Thus, the total cost of solving (27) to global optimality through this reformulation is bounded by the cost of solving $2n$ linear programs, the largest of which has $\mathcal{O}(n)$ variables and constraints, and the smallest of which has $\mathcal{O}(1)$ variables and constraints.

References

1. Ben-Tal, A., den Hertog, D., Vial, J.P.: Deriving robust counterparts of nonlinear uncertain inequalities. Mathematical Programming **149**(1), 265–299 (2015). doi:10.1007/s10107-014-0750-8
2. Ben-Tal, A., El Ghaoui, L., Nemirovski, A.: Robust Optimization. Princeton University Press (2009)
3. Ben-Tal, A., Hazan, E., Koren, T., Mannor, S.: Oracle-based robust optimization via online learning. Operations Research **63**(3), 628–638 (2015). doi:10.1287/opre.2015.1374
4. Bertsimas, D., Brown, D., Caramanis, C.: Theory and applications of robust optimization. SIAM Review **53**(3), 464–501 (2011). doi:10.1137/080734510
5. Bertsimas, D., Dunning, I., Lubin, M.: Reformulation versus cutting-planes for robust optimization. Computational Management Science **13**(2), 195–217 (2016). doi:10.1007/s10287-015-0236-z
6. Bertsimas, D., Nohadani, O.: Robust optimization with simulated annealing. Journal of Global Optimization **48**(2), 323–334 (2010). doi:10.1007/s10898-009-9496-x
7. Bertsimas, D., Nohadani, O., Teo, K.M.: Robust optimization for unconstrained simulation-based problems. Operations Research **58**(1), 161–178 (2010). doi:10.1287/opre.1090.0715
8. Calafiore, G., Campi, M.: Uncertain convex programs: randomized solutions and confidence levels. Mathematical Programming **102**(1), 25–46 (2005). doi:10.1007/s10107-003-0499-y
9. Cheney, E.W., Goldstein, A.A.: Newton's method for convex programming and Tchebycheff approximation. Numerische Mathematik **1**, 253–268 (1959). doi:10.1007/bf01386389
10. Ciccazzo, A., Latorre, V., Liuzzi, G., Lucidi, S., Rinaldi, F.: Derivative-free robust optimization for circuit design. Journal of Optimization Theory and Applications **164**(3), 842–861 (2015). doi:10.1007/s10957-013-0441-2
11. Conn, A.R., Gould, N.I.M., Toint, P.L.: Trust-Region Methods. Society for Industrial and Applied Mathematics (2000)
12. Conn, A.R., Scheinberg, K., Vicente, L.N.: Introduction to Derivative-Free Optimization. Society for Industrial and Applied Mathematics (2009)
13. Conn, A.R., Vicente, L.N.: Bilevel derivative-free optimization and its application to robust optimization. Optimization Methods and Software **27**(3), 561–577 (2012). doi:10.1080/10556788.2010.547579
14. Diehl, M., Bock, H.G., Kostina, E.: An approximation technique for robust nonlinear optimization. Mathematical Programming **107**(1-2), 213–230 (2006). doi:10.1007/s10107-005-0685-1
15. Duran, M.A., Grossmann, I.E.: An outer-approximation algorithm for a class of mixed-integer nonlinear programs. Mathematical Programming **36**(3), 307–339 (1986). doi:10.1007/BF02592064
16. Fletcher, R., Leyffer, S.: Solving mixed integer nonlinear programs by outer approximation. Mathematical Programming **66**(1), 327–349 (1994). doi:10.1007/BF01581153
17. Hettich, R., Kortanek, K.O.: Semi-infinite programming: Theory, methods, and applications. SIAM Review **35**(3), 380–429 (1993). doi:10.1137/1035089
18. Kelley, Jr., J.E.: The cutting-plane method for solving convex programs. Journal of the Society for Industrial and Applied Mathematics **8**(4), 703–712 (1960). doi:10.1137/0108053
19. Khan, K., Larson, J., Wild, S.M.: Manifold sampling for nonconvex optimization of piecewise linear compositions. Preprint ANL/MCS-P8001-0817, Argonne National Laboratory, MCS Division (2017). URL `http://www.mcs.anl.gov/papers/P8001-0817.pdf`
20. Larson, J., Menickelly, M., Wild, S.M.: Manifold sampling for L1 nonconvex optimization. SIAM Journal on Optimization **26**(4), 2540–2563 (2016). doi:10.1137/15M1042097
21. Moré, J.J., Wild, S.M.: Benchmarking derivative-free optimization algorithms. SIAM Journal on Optimization **20**(1), 172–191 (2009). doi:10.1137/080724083
22. Polak, E.: Optimization. Springer New York (1997). doi:10.1007/978-1-4612-0663-7
23. Postek, K., den Hertog, D., Melenberg, B.: Computationally tractable counterparts of distributionally robust constraints on risk measures. SIAM Review **58**(4), 603–650 (2016). doi:10.1137/151005221
24. Wild, S.M., Regis, R.G., Shoemaker, C.A.: ORBIT: Optimization by radial basis function interpolation in trust-regions. SIAM Journal on Scientific Computing **30**(6), 3197–3219 (2008). doi:10.1137/070691814

25. Wild, S.M., Shoemaker, C.A.: Global convergence of radial basis function trust-region algorithms for derivative-free optimization. SIAM Review **55**(2), 349–371 (2013). doi:10.1137/120902434