Scalable Projection-Free Optimization Methods via MultiRadial Duality Theory

Thabo Samakhoana* Benjamin Grimmer†

Abstract

Recent works have developed new projection-free first-order methods based on utilizing linesearches and normal vector computations to maintain feasibility. These oracles can be cheaper than orthogonal projection or linear optimization subroutines but have the drawback of requiring a known strictly feasible point to do these linesearches with respect to. In this work, we develop new theory and algorithms which can operate using these cheaper linesearches while only requiring knowledge of points strictly satisfying each constraint separately. Convergence theory for several resulting "multiradial" gradient methods is established. We also provide preliminary numerics showing performance is essentially independent of how one selects the reference points for synthetic quadratically constrained quadratic programs.

1 Introduction

Recently, several works [1–9] have proposed new projection-free first-order methods based on often cheap linesearches and normal vector computations with the feasible region. Such methods offer potential advantages in terms of their scalability over projected methods and conditional gradient/Frank-Wolfe-type methods as reliances on quadratic or linear optimization oracles as subroutines are avoided. Prior works based on such potentially cheaper linesearches have required knowledge of a "good enough" strictly feasible point to use as a reference. In the line of work by Grimmer [5, 6, 10], these methods are called radial methods as linesearches based at the origin amount to searching along rays at each iteration. In this work, we circumvent the previous reliance on a known "good enough" strictly feasible point by developing a new family of "MultiRadial Methods". These methods instead rely on a collection of reference points, each only required to be feasible to one component of the problem's constraints.

Our primary interest is in the development of methods for maximization problems

$$p^* = \begin{cases} \max\ f(x) \\ \text{s.t. } x \in S_j \quad \text{for all } j = 1, \dots, m \end{cases} \tag{1.1}$$

with concave objective function $f : \mathcal{E} \to \mathbb{R} \cup \{-\infty\}$ and closed convex constraint sets $S_j \subseteq \mathcal{E}$ for some finite dimensional Euclidean space $\mathcal{E}$. No assumptions like Lipschitz continuity of f are made. We focus on the development of first-order methods where f can be accessed through its function value, its (sup)gradients, and one-dimensional linesearches. Mirroring these three operations, we will only assume access to the sets S_j via checking membership, its normal vectors, and one-dimensional linesearches.

*Johns Hopkins University, Department of Applied Mathematics and Statistics, tsamakh1@jhu.edu
†Johns Hopkins University, Department of Applied Mathematics and Statistics, grimmer@jhu.edu

Alternative commonly utilized oracle models for the constraint sets S_j can incur higher per-iteration computational costs. Orthogonal projections, commonly used in projected gradient methods, require quadratic optimization over each S_j (or worse $\cap S_j$), which requires S_j to be sufficiently simple this can be done in closed-form (or quickly approximated). Frank-Wolfe-type methods only require linear optimization at each iteration, which is often cheaper than projections but may still be prohibitive. Interior point-type methods are applicable when a self-concordant barrier function for each S_j is available but require linear systems solves based at each iteration.

Lagrangian-type methods apply when the constraints take the functional form of $S_j = \{x \mid g_j(x) \leq 0\}$, relying on first-order oracles for and the structure of each g_j. If each g_j is convex but nonsmooth, a range of subgradient-type methods can be applied [11, 12]. If each g_j is smooth, nearly optimal accelerated methods have been recently developed by Zhang and Lan [13]. An important distinction should be drawn between using first-order evaluations of functional constraints g_j and our model of linesearches and normal vectors of S_j. Our oracle is independent of how one represents the set S_j. In contrast, the above referenced methods for functionally constrained problems may require careful preprocessing of constraints to perform well, as, for example, replacing $g_j(x) \leq 0$ with any positive rescaling $\lambda g_j(x) \leq 0$ will change their algorithm's trajectory.

Here we develop algorithms that access each constraint set S_j by linesearches and normal vector computations. As linesearches, given some $e_j \in \operatorname{int} S_j$ and $x \notin S_j$, we assume one can find the unique point on the boundary of S_j between e_j and x. Even if this cannot be done in closed form, given a membership oracle for S_j, bisection or a similar rootfinding procedure could be used to reach a machine precision solution. Once a boundary point is produced, we assume a normal vector can be computed, mirroring the role of computing (sub)gradients of the objective. These two operations correspond to function evaluation and subgradient evaluation of the gauge of S_j with respect to e_j, defined as

$$\gamma_{S_j,e_j}(x) = \inf\left\{v > 0 \mid e_j + \frac{x - e_j}{v} \in S_j\right\} .$$

(A formal introduction and discussion of gauges is deferred to Section 2.1.)

These two oracles are often much cheaper (and hence lead to more scalable algorithms) than common alternatives. For example, consider any ellipsoidal constraint $S_j = \{x \mid \|A_j x - b_j\|_2 \leq 1\}$. Here our assumed linesearch and normal vector can be cheaply computed with closed forms: the one-dimensional linesearch is directly given by the quadratic formula and a normal vector follows from one matrix multiplication with $A_j^T A_j$. In contrast, linear optimization, projections, and interior point method steps on ellipsoids all require at least solving a linear system.

A family of projection-free algorithms only utilizing these cheaper oracles was first developed by Renegar [3, 4]. We introduce these ideas following their more general development as "radial algorithms" of Grimmer [6, 10]. These methods reformulate (1.1) as the equivalent radially dual problem[1]

$$\min_y \max_j \{f^{\Gamma,e}(y), \gamma_{S_j,e}(y)\} \tag{1.2}$$

provided $f(e) > 0$ and $e \in \operatorname{int} \cap S_j$. Here $\gamma_{S,e}$ is the gauge of S_j with respect to e and $f^{\Gamma,e}$ is a nonlinear transformation of f (again see Section 2.1 for formal definitions). This reformulation is quite amenable to the application of first-order methods since (i) it is unconstrained minimization, facilitating the use of projection-free methods, (ii) it only interacts with the constraints S_j through their gauges, enabling the use of often cheaper oracles, and (iii) it is uniformly Lipschitz continuous, removing the need to assume such structure. However, the applicability of prior radial algorithms based on solving (1.2) is limited by the required knowledge of a common strictly feasible point e.

[1]Note this radial dual is fundamentally different from the similarly named gauge dual of Freund [14] as knowledge of oracles for related conjugate functions and polar sets are avoided in the radial dual formulation.

Indeed, the Lipschitz continuity of (1.2) depends on how interior e is to $\cap S_j$. So, a "good" reference point is very much needed for prior methods to be effective.

Our Contributions The primary contribution of this work is generalizing the duality between the primal problem (1.1) and radially dual problem (1.2) preserving the benefits (i)-(iii) above while avoiding any usage of a common point e. Instead, we consider the MultiRadially Dual problem

$$\min_y \max_j \{f^{\Gamma,e_0}(y), \gamma_{S_j,e_j}(y)\} \tag{1.3}$$

which only relies on separate points e_0 with $f(e_0) > 0$ and $e_j \in \text{int } S_j$ for each constraint. More generally, we develop theory relating (1.1) to any problem of the form $\min_y \max\{f^{\Gamma,e_0}(y), \varphi(y)\}$ where $\varphi : \mathcal{E} \to \mathbb{R}$ is a convex function "identifying" the feasible region $\cap S_j = \{x \in \mathcal{E} \mid \varphi(x) \leq 1\}$.

1. **MultiRadial Duality Theory** We develop theory relating the optimal solutions of the primal problem (1.1) to those of (1.3). Our Theorems 3.3 and 3.4 provide direct, algorithmically useful bounds relating the primal and multiradial dual optimal values, controlled by a natural geometric condition number. In the special case where $p^* = 1$, these bounds become tight and our Theorem 3.1 shows both problems have exactly the same solution sets.

2. **MultiRadial Methods** Based on this theory, we design and analyze new scalable, projection-free "MultiRadial Methods". For nonLipschitz nonsmooth convex optimization, our Corollary 4.1 guarantees a MultiRadial Subgradient Method converges at the optimal $O(1/\varepsilon^2)$ rate up to a log term, with each iteration computing at most one subgradient of f or one normal vector of a constraint. When the objective and constraint sets are smooth, our Corollaries 4.2 and 4.3 show accelerated MultiRadial Smoothing and Generalized Gradient Methods converge at rates $O(1/\varepsilon)$ and $O(1/\sqrt{\varepsilon})$ up to a log term, where the latter relies on more expensive per-iteration computations with respect to m.

Example - Convex Quadratically Constrained Quadratic Programming (QCQPs) Throughout this work, we periodically utilize quadratic optimization problems as a concrete, classic model to illustrate results. In particular, consider a convex QCQP

$$p^* = \begin{cases} \max & f_0(x) := r_0 - q_0^T x - \frac{1}{2} x^T P_0 x \\ \text{s.t.} & f_j(x) := r_j - q_j^T x - \frac{1}{2} x^T P_j x \geq 0 \quad \forall j = 1 \dots m \ . \end{cases} \tag{1.4}$$

for any positive semidefinite matrices P_j and $p^* > 0$.

For convex QCQPs, one natural selection for e_0 is the maximizer of the objective $f(x)$, given by solving $P_0 e + q_0 = 0$. Similarly, a natural selection of e_j would be any solution of $P_j e + q_j = 0$. Our approach applies for any selection of e_j's with $f_j(e_j) > 0$. In Section 5, we numerically observe that the typical numerical performance of our MultiRadial Methods tends to be independent of the choice of centers e_j. Consequently, it may suffice to cheaply approximate a solution of $P_j e + q_j = 0$.

Supposing each P_j is positive definite, these selections correspond to $e_j = -P_j^{-1} q_j$ for $j = 0, \dots, m$. Then the multiradial dual problem (1.3) of (1.4) takes the form

$$\min_y \max_{j=1\dots m} \left\{ \frac{1 + \sqrt{1 + 2f_0(e_0)(y-e_0)^T P_0 (y-e_0)}}{2f_0(e_0)}, \sqrt{\frac{(y-e_j)^T P_j (y-e_j)}{2f_j(e_j)}} \right\} \ . \tag{1.5}$$

More generally, for any positive semidefinite P_j and any selection of e_j with $f_j(e_j) > 0$, the multiradial dual problem (1.3) of (1.4) remains describable in closed form as

$$\min_y \max_{j=1\dots m} \left\{ \frac{1 - \nabla f_0(e_0)^T(y-e_0) + \sqrt{(1-\nabla f_0(e_0)^T(y-e_0))^2 + 2f_0(e_0)(y-e_0)^T P_0(y-e_0)}}{2f_0(e_0)}, \right.$$
$$\left. \frac{-\nabla f_j(e_j)^T(y-e_j) + \sqrt{(\nabla f_j(e_j)^T(y-e_j))^2 + 2f_j(e_j)(y-e_j)^T P_j(y-e_j)}}{2f_j(e_j)} \right\}. \quad (1.6)$$

In either case, each component of the objective and its gradient can be computed via one matrix-vector multiplication. In this sense, we claim the resulting multiradial first-order methods are "scalable" as many existing alternatives require at least a linear system solve each iteration. The development of method's only relying on matrix-vector multiplication has been a recent trend in linear programming [15–18] and quadratic programming [19].

Outline Section 2 introduces needed preliminaries. Our theory in Section 3 relates our unconstrained "multiradial" reformulations to the original problem and discusses immediate algorithmic consequences. Subsequently, in Section 4, we develop a parameter-free method based on approximately solving (rescalings of) these multiradial problems. Preliminary numerical results are presented in Section 5 for QCQPs, validating our theory and highlighting one area where performance scales better than our theory predicts.

2 Preliminaries

Our notations follow those of the initial development of radial duality [6,10], specialized to the convex settings considered here. We consider any finite-dimensional Euclidean space $\mathcal{E}$ with a norm $\|\cdot\|$ induced by an inner product $\langle\cdot,\cdot\rangle$. To apply previous radial theory, we restrict to consider objective functions with values in the (extended) positive reals, which we denote by $\overline{\mathbb{R}}_{++} = \mathbb{R}_{++} \cup \{0, \infty\}$. Here, $\mathbb{R}_{++}$ is the set of positive real numbers and $0, \infty$ should be interpreted as the limit points of $\mathbb{R}_{++}$, playing a similar role to $\pm\infty$ for the real numbers.

Throughout, we will primarily consider extended positive valued functions $f : \mathcal{E} \to \mathbb{R}_{++} \cup \{0, \infty\}$. We claim this restriction is minor: for any real-valued objective $\tilde{f} \colon \mathcal{E} \to \mathbb{R} \cup \{\pm\infty\}$ to be maximized, one can equivalently maximize the extended positive valued function $f(x) := \max\{\tilde{f}(x) - \tilde{f}(x_0) + 1, 0\}$ when given any $x_0 \in \mathcal{E}$ with $f(x_0) \in \mathbb{R}$. For any extended real-valued function f, its effective domain, epigraph, and hypograph are

$$\begin{aligned} \text{dom } f &:= \{x \in \mathcal{E} \mid f(x) \in \mathbb{R}_{++}\} \\ \text{epi } f &:= \{(x,u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \le u\} \\ \text{hypo } f &:= \{(x,u) \in \mathcal{E} \times \mathbb{R}_{++} \mid f(x) \ge u\}, \end{aligned}$$

respectively. We denote the closure of dom f by S_0. A function $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is concave (convex) if hypo f (epi f) is convex. We say f is upper (lower) semicontinuous at $x \in \mathcal{E}$ if $\limsup_{x' \to x} f(x') = f(x)$ ($\liminf_{x' \to x} f(x') = f(x)$) and say f is globally upper (lower) semicontinuous if this holds for all $x \in \mathcal{E}$. We abbreviate upper (lower) semicontinuity as u.s.c. (l.s.c.) at times.

Normals, Subdifferentials, and Smoothness. The inner product on $\mathcal{E}$ induces one on $\mathcal{E} \times \mathbb{R}$ defined by $\langle (x,u), (x',u') \rangle := \langle x, x' \rangle + u \cdot u'$. We use the same notation for both inner products as

it will be clear from context which is being used. We say that a vector ξ is normal to a set S at x if $\langle \xi, x' - x \rangle \leq 0$ for all $x' \in S$. The set of all normal vectors to S at x is denoted by $N_S(x)$. A vector $\zeta \in \mathcal{E}$ is a subgradient of convex function f at $x \in \mathcal{E}$ if $(\zeta, -1) \in N_{\text{epi } f}((x, f(x))$. The set of all subgradients of f at x is denoted by $\partial f(x)$ and referred to as the subdifferential of f at x. We say $\zeta \in \mathcal{E}$ is a supgradient of a concave function f at $x \in \mathcal{E}$ if $(-\zeta, 1) \in N_{\text{hypo } f}((x, f(x))$. If f is continuously differentiable, these differentials are exactly the singleton $\{\nabla f(x)\}$.

We say a function $f : \mathcal{E} \to \mathbb{R}$ is M-Lipschitz continuous if $|f(x) - f(y)| \leq M\|x - y\|$ for all $x, y \in \mathcal{E}$ and a continuously differentiable function f is L-smooth if its gradient is L-Lipschitz continuous on its domain. We say a set S is β-smooth if any two unit length normal vectors $\xi_i \in N_S(x_i)$ for $i \in \{1, 2\}$ satisfy $\|\xi_1 - \xi_2\| \leq \beta\|x_1 - x_2\|$. A more detailed discussion on smooth sets is given in [9].

2.1 Minkowski Gauges and Radial Reformulations

For any set $S \subseteq \mathcal{E}$, we define its gauge with respect to some $e \in S$ as

$$\gamma_{S,e}(x) := \inf \left\{ v > 0 \mid e + \frac{x - e}{v} \in S \right\} . \tag{2.1}$$

When $e = 0$, this is the Minkowski gauge, denoted by $\gamma_S(y) = \inf\{v > 0 \mid e/v \in S\}$. Otherwise, $\gamma_{S,e}$ can be viewed as a translation of the Minkowski gauge γ_{S-e}. Note if S is convex and $e \in \text{int } S$, then $\gamma_{S,e}$ is convex, continuous and finite everywhere.

This gauge of a set has a close relationship to the following indicator function. Namely, consider the nonstandard indicator function $\hat{\iota}_S : \mathcal{E} \to \{0, \infty\}$ defined as

$$\hat{\iota}_S(x) := \begin{cases} +\infty & \text{if } x \in S \\ 0 & \text{otherwise .} \end{cases} \tag{2.2}$$

To relate these functions, observe that the hypograph of this indicator has a bijection to the epigraph of the gauge of a closed convex S with respect to any $e \in \text{int } S$ of

$$\Gamma_e(x, u) := \left(e + \frac{x - e}{u}, \frac{1}{u} \right) . \tag{2.3}$$

Namely,

$$\text{hypo } \hat{\iota}_S = \Gamma_e(\text{epi } \gamma_{S,e}). \tag{2.4}$$

This "radial transformation" Γ_e was introduced in [10], fixing $e = 0$.

The epigraph-hypograph bijection (2.4) motivates the following radial function transformation of a generic function $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ with $e \in \mathcal{E}$ as[2]

$$f^{\Gamma,e}(y) := \sup \{ v > 0 \mid (y, v) \in \Gamma_e(\text{epi } f) \} . \tag{2.5}$$

Intuitively, one can view $f^{\Gamma,e}$ as the smallest function whose hypograph contains $\Gamma_e(\text{epi } f)$. When $f = \gamma_{S,e}$ for a closed convex set S with $e \in \text{int } S$, this radial transformation exactly turns gauges into indicator functions $\hat{\iota}_S^{\Gamma,e} = \gamma_{S,e}$. Moreover, one can verify the reverse holds as well, $\hat{\iota}_S = \gamma_{S,e}^{\Gamma,e}$. So this transformation provides a bijection between indicator and gauge functions.

Expanding the definitions of Γ_e and epi f, one has $f^{\Gamma,e}(y) = \sup \left\{ v > 0 \mid v f(e + \frac{y-e}{v}) \leq 1 \right\}$. When $e = 0$, we ease notation, writing $f^{\Gamma} = f^{\Gamma,0}$. From this, it becomes clear that $f^{\Gamma,e} = (f \circ \mathfrak{t}_e)^{\Gamma} \circ \mathfrak{t}_{-e}$

[2] If the set on the right of (2.5) is empty, we set $f^{\Gamma,e}(y) = 0$ rather than $-\infty$ to ensure the transformed function also maps into the extended positive reals.

where $\mathfrak{t}_e(y) = e + y$ denotes a translation by e, and so this radial transformation is just a translation of those proposed by Grimmer [6, 10]. In the following, we summarize their results relating f to f^Γ and $(f^\Gamma)^\Gamma$, emphasizing that f^Γ can be replaced with $f^{\Gamma,e}$ for $e \in \mathcal{E}$ by the simple translation argument noted above. For more exposition, we refer the reader to the relevant parts of [10] and [6].

The duality between indicators and gauges of convex sets carries over more generally to a wide range of (potentially nonconvex) functions. In particular, we say f is upper radial with respect to e if the translated perspective function $f^{p,e}(x, v) = vf(e + \frac{x-e}{v})$ is upper semicontinuous and nondecreasing in $v > 0$ for all fixed $x \in \mathcal{E}$. Theorem 1 of [10] establishes that this condition exactly characterizes when the radial function transformation is dual: For any $e \in \mathcal{E}$,

$$(f^{\Gamma,e})^{\Gamma,e} = f \text{ if and only if } f \text{ is upper radial with respect to } e \ . \tag{2.6}$$

The condition that $f^{p,e}(x, \cdot)$ is nondecreasing for all $x \in \mathcal{E}$ is equivalent to hypo f being star-convex with respect to $(e, 0)$, cf. [10, Lemma 1]. This duality between functions extends to give a duality between optimization problems as for any such objective: Proposition 24 of [10] ensures

$$(\operatorname{argmax}\ f) \times \{\max f\} = \Gamma_e \left(\left(\operatorname{argmin}\ f^{\Gamma,e} \right) \times \left\{ f^{\Gamma,e} \right\} \right) \ . \tag{2.7}$$

Structural Properties of Gauges and Radial Reformulations. This work is primarily concerned with concave objective functions f being maximized over convex sets S_j, for which the above star-convexity condition is easily verified. In this case, we can ensure a strengthened version of upper radiality holds: when f is upper radial with respect to e and $f^{p,e}(x, \cdot)$ is strictly increasing on dom $f^{p,e}(x, \cdot) := \{v > 0 \mid f^{p,e}(x, v) \in \mathbb{R}_{++}\}$ for every $x \in \mathcal{E}$, we say f strictly upper radial with respect to e. Then, it follows that all functions and sets considered here are well behaved as

$$f \text{ is concave and u.s.c. } \implies f \text{ is strictly upper radial w.r.t. any } e \in \operatorname{int} \operatorname{dom} f \tag{2.8}$$

$$S \text{ is convex and closed } \implies \hat{\iota}_S \text{ is strictly upper radial w.r.t. any } e \in \operatorname{int} S \ . \tag{2.9}$$

Given a bound on how interior e is to the domain of f (or to the constraint set S), we can further guarantee the radial transformation (or gauge) with respect to e is well behaved, i.e., convex and uniformly Lipschitz continuous. Denote the interior radius of S with respect to e and diameter by

$$R_e(S) := \inf \{\|x - e\| \mid x \notin S\}$$
$$D(S) := \sup \{\|x - y\| \mid x, y \in S\}$$

Then [10, Proposition 17] and [6, Proposition 1, Lemma 1] ensure the following

$$f \text{ is concave, u.s.c., and } R_e(S_0) > 0 \implies f^{\Gamma,e} \text{ is convex and } 1/R_e(S_0)\text{-Lipschitz} \ , \tag{2.10}$$

$$S \text{ is convex, closed, and } R_e(S) > 0 \implies \gamma_{S,e} \text{ is convex and } 1/R_e(S)\text{-Lipschitz} \tag{2.11}$$

where $S_0 = \operatorname{cl} \operatorname{dom} f$. Hence, provided "good" interior points to the domain of f and each constraint are known, their transformations will be well-behaved and conditioned[3]. Moreover, when f is L-smooth or S is β-smooth, this structure is preserved. Namely [6, Proposion 2] ensures for twice continuously differentiable f with bounded domain, $f^{\Gamma,e}$ is $O(L)$-smooth and [9, Theorem 3.2] ensures for β-smooth, compact S, $\gamma^2_{S,e}$ is $O(\beta)$-smooth. Both big-O statements above suppress constants depending on the geometric radius and diameter quantities above.

[3]In the nonconvex development of these radial transformations of [6], these R constants are generalized to measure how star-convex the given function's hypograph is.

Finally, we note three calculus/computational results of interest to our development. The family of upper and strictly upper radial functions is closed under many common operations, see [10, Propositions 12 and 13]: If f is (strictly) upper radial with respect to e, then so is λf for all $\lambda > 0$ and

$$(\lambda f)^{\Gamma,e} = \frac{1}{\lambda} f^{\Gamma,e} \circ \mathfrak{t}_e \circ \lambda \mathfrak{t}_{-e} \ . \tag{2.12}$$

If f_1, f_2 are both (strictly) upper radial with respect to e, then so is $\min\{f_1, f_2\}$ and

$$(\min\{f_1, f_2\})^{\Gamma,e} = \max\left\{f_1^{\Gamma,e}, f_2^{\Gamma,e}\right\}. \tag{2.13}$$

For any f that is strictly upper radial with respect to some e, the subgradients of $f^{\Gamma,e}$ are easily computed from those of f as [10, Proposition 19] ensures

$$\partial f^{\Gamma,e}(y) = \left\{ \frac{\zeta}{\langle (\zeta,\delta), (x-e,u)\rangle} \mid (\zeta,\delta) \in N_{\text{hypo } f}((x,u)), \langle (\zeta,\delta), (x-e,u)\rangle > 0 \right\} \tag{2.14}$$

where $(x,u) = \Gamma_e((y, f^{\Gamma,e}(y)))$.

2.2 First-Order Methods Minimizing Finite Maximums

Instead of directly solving the primal problem (1.1), our proposed MultiRadial Methods will solve (a sequence of) unconstrained convex minimization problems of the form (1.3). These reformulations will always be minimizing a finite maximum of convex functions:

$$h_\star = \min_x \max\{h_0(x), \ldots, h_m(x)\} \ . \tag{2.15}$$

Let $h(x) = \max\{h_0(x), \ldots, h_m(x)\}$ denote the whole objective being minimized. Depending on the structure of f and S_j in (1.1), the multiradial dual will have components h_j that are either Lipschitz or smooth. Below we review three well-known families of first-order methods capable of minimizing such objectives: first, the subgradient method for nonsmooth settings, and then accelerated smoothing and generalized gradient methods for smooth settings with large or small values of m, respectively.

Each first-order method `fom` considered maintains a sequence of iterates y_i defined by two (simple) procedures for initializing/restarting itself and for taking one step. We denote the initialization process by $y_0 = \texttt{fom.initialize}(x, \varepsilon, h)$, where $x \in \mathcal{E}$ is an initial solution, $\varepsilon > 0$ is a target accuracy, and h is the objective to minimize. For momentum methods, this procedure may involve initializing auxiliary variable sequences as well. We denote taking one step of `fom` by $y_{i+1} = \texttt{fom.step}(y_i, \varepsilon, h)$, although auxiliary variable sequences may be updated as well. The considered methods all have convergence guarantees of the following form: If $\|y_0 - y^*\| \le D$ for some minimizer y^* of h, then

$$\text{Some } i \le K_{\texttt{fom}}(D, \varepsilon, h) \text{ has } h(y_i) - h(y^*) \le \varepsilon \ . \tag{2.16}$$

The Subgradient Method The subgradient method, dubbed `subgrad`, initializes simply with $y_0 = x_0$ and iterates

$$y_{i+1} = y_i - \varepsilon g_i / \|g_i\|^2, \qquad g_i \in \partial h(y_i) \ . \tag{2.17}$$

Note a subgradient of $h(x_k)$ can be computed as any subgradient of some $h_j(x_k)$ attaining the finite maximum. Provided each h_j is convex and M-Lipschitz, which implies h is convex and M-Lipschitz, the convergence of this method is well studied, having $K_{\texttt{subgrad}}(D, \varepsilon, h) = M^2 D^2 / \varepsilon^2$.

The (Accelerated) Smoothing Method Supposing instead that each h_j is L-smooth and M-Lipschitz, one can utilize the smoothing techniques of [20,21]. Given a target accuracy $\varepsilon > 0$, one can approximate h by $h_\theta(y) = \theta \log\left(\sum_{j=0}^m \exp\left(\frac{h_j(y)}{\theta}\right)\right)$ for $\theta = \frac{\varepsilon}{2\log(m+1)}$. One can verify h_θ has $|h_\theta - h| \leq \varepsilon/2$ and is $L_\theta = L + \frac{M^2}{\theta}$-smooth. Then one can apply any accelerated gradient method to minimize h_θ. For example, Nesterov's accelerated method initialized with $z_0 = y_0, t_0 = \frac{-1+\sqrt{5}}{2}$ iterates

$$\begin{cases} y_{i+1} = z_i - \frac{1}{L_\theta}\nabla h_\theta(z_i) \\ z_{i+1} = y_{i+1} + \beta_i(y_{i+1} - y_i) \end{cases} \tag{2.18}$$

where $t_{i+1}^2 = (1-t_i)t_i^2$ and $\beta_i = t_i(1-t_i)/(t_i^2 + t_{i+1})$. We denote this method by `smooth`. Noting any $\varepsilon/2$-minimizer of h_θ is an ε-minimizer of h, the accelerated convergence of $2\sqrt{L_\theta D^2/\varepsilon}$ in [22, Theorem 2.2.3] gives a guarantee of the form (2.16)

$$K_{\texttt{smooth}}(D, \varepsilon, h) = 2\sqrt{\frac{2LD^2}{\varepsilon} + \frac{4M^2D^2\log(m+1)}{\varepsilon^2}}\ .$$

In our numerics, we will instead use the Universal Fast Gradient Method (UFGM) of Nesterov [23], which avoids requiring knowledge of L_θ.

The (Accelerated) Generalized Gradient Method If, in addition to being L-smooth, the number of terms in the finite maximum m is relatively small, one can utilize the generalized gradient method as outlined in [22]. This method works by utilizing the generalized gradient mapping defined as

$$\mathcal{G}(y, \alpha) = \frac{1}{\alpha}\operatorname*{argmin}_y \left\{ \max_{j=0,\dots,m} \left\{ h_j(y) + g_j^T(y' - y) \right\} + \frac{1}{2\alpha}\|y' - y\|^2 \right\}, \qquad g_j \in \partial h_j(y),$$

and then applying any accelerated method with $\mathcal{G}(y, \alpha)$ replacing the gradient, which we dub `genGrad`. Computing $\mathcal{G}(y, \alpha)$ corresponds to solving a quadratic program of dimension $m+1$. This limits the applicability of such methods to settings where this can be efficiently calculated, primarily being useful when m is small. Theorem 2.3.5 of [22] ensures this method has a convergence guarantee of the form (2.16) with $K_{\texttt{genGrad}}(D, \varepsilon, h) = 2\sqrt{LD^2/\varepsilon}$.

3 MultiRadial Theory and Idealized Methods

We begin by developing our multiradial duality theory relating generic constrained maximization problems (1.1) to the unconstrained multiradially dual problem (1.3). Throughout, we will discuss immediate algorithmic implications by analyzing resulting simple multiradial algorithms. In the following section, we will propose and analyze a more practical parameter-free multiradial method.

First, we introduce some notations to describe the primal and (multi)radial dual objectives of (1.1) and (1.3). Let $\mathcal{S} := \bigcap_{j=1}^m S_j$ denote the primal feasible region and $\Psi : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ denote the primal function

$$\Psi(x) := \min\left\{f(x), \hat{\iota}_{\mathcal{S}}(x)\right\}\ . \tag{3.1}$$

Maximizing $\Psi(x)$ is exactly the original primal problem (1.1) provided some $x \in \mathcal{S} \cap \operatorname{dom} f$ exists, so $p^* := \max_{x\in\mathcal{S}} f(x) = \max_{x\in\mathcal{E}} \Psi(x)$. For any $e \in \operatorname{int}(\mathcal{S} \cap \operatorname{dom} f)$,

$$\Psi^{\Gamma,e} = \max\left\{f^{\Gamma,e}, \gamma_{\mathcal{S},e}\right\} \tag{3.2}$$

by equation (2.13) and the fact that $\gamma_{\mathcal{S},e} = \hat{\imath}_{\mathcal{S}}^{\Gamma,e}$. Thus the following duality relation holds

$$(\text{argmax } \Psi) \times \{\max \Psi\} = \Gamma_e \left(\left(\text{argmin } \Psi^{\Gamma,e} \right) \times \left\{ \min \Psi^{\Gamma,e} \right\} \right) \tag{3.3}$$

by (2.7). Requiring a point e interior to every constraint is a notable limitation to the design of algorithms based on this relation. We address this by relaxing the dual objective function $\Psi^{\Gamma,e}$. We instead consider the following dual function

$$\Phi(y) = \max\{f^{\Gamma,e_0}(y), \varphi_{\mathcal{S}}(y)\} \tag{3.4}$$

where $\varphi_{\mathcal{S}} : \mathcal{E} \to \mathbb{R}$ is a l.s.c. convex function satisfying $\text{int } \mathcal{S} = \{x \in \mathcal{E} \mid \varphi_{\mathcal{S}}(x) < 1\}$. We call any such $\varphi_{\mathcal{S}}$ a *convex identifier* of $\mathcal{S}$. Based on equation (3.2), a natural choice for an identifier is $\varphi_{\mathcal{S}} = \max\{\gamma_{S_1,e_1}, \ldots, \gamma_{S_m,e_m}\}$ where $e_j \in \text{int } S_j$ for all j. This particular $\varphi_{\mathcal{S}}$ enables us to replace $e \in \text{int } (\mathcal{S} \cap \text{dom } f)$ with separate reference points for each functional component of the primal objective (3.1). For this reason, we call Φ in (3.4) the multiradial dual function. We will at times refer to $\max\{\gamma_{S_1,e_1}, \ldots, \gamma_{S_m,e_m}\}$ as the canonical $\varphi_{\mathcal{S}}$ and we encourage the reader to keep it as a concrete example of a convex identifier.

The primal problem and (multiradial) dual problem are then given by

$$p^* := \max_{x \in \mathcal{E}} \Psi(x) \tag{3.5}$$

$$d^* := \min_{y \in \mathcal{E}} \Phi(y) \ . \tag{3.6}$$

We will show that, under suitable assumptions, Φ is indeed an appropriate replacement to the radial dual $\Psi^{\Gamma,e}$ given by using a single reference point. A condition analogous to (3.3) is derived in Theorem 3.1 in a restricted case, with general relationships being given in Theorems 3.3 and 3.4. Note that with the canonical $\varphi_{\mathcal{S}}$, the multiradial dual problem is an unconstrained, convex, uniformly Lipschitz minimization problem (and thus remains amenable to the direct application of many first-order methods).

Our theory relies on four assumptions, ensuring (1.1) is concave maximization with a maximizer and a Slater point, and that $\varphi_{\mathcal{S}}$ and f^{Γ,e_0} are well defined.

Assumption A. *f is concave and u.s.c. with bounded zero super-level set*

$$D_0 := D(S_0) < \infty.$$

Assumption B. *The constraint sets $S_1, \ldots, S_m$ are convex and closed.*

Assumption C. *A convex identifier $\varphi_{\mathcal{S}}$ is known and a point $e_0 \in \text{int } S_0$ is known with*

$$R_0 := R_{e_0}(S_0) > 0.$$

Assumption D. *There exists $x^* \in \mathcal{S}$ with $f(x^*) = p^* > 0$ and $x_{SL} \in \text{int } \mathcal{S} \cap \text{dom } f$ such that*

$$\eta := (1 - \gamma_{S_0,e_0}(x^*))(1 - \varphi_{\mathcal{S}}(x_{SL})) > 0.$$

A few notes on these conditions. Firstly, under Assumptions A and B, Assumption C is satisfied if points $e_j \in \text{int } S_j$ are known for each $j = 0, 1, \ldots, m$. In this case, with $R := \min\{R_{e_j}(S_j) \mid j = 0, 1, \ldots, m\}$, Φ with the canonical $\varphi_{\mathcal{S}}$ is $1/R$-Lipschitz continuous. Note the multiradial reformulation Φ can have a better Lipschitz constant than the radial dual (1.2) relying on knowing a single $e \in \text{int } \cap_{j=0}^m S_m$ which is $1/R_e(\cap_{j=0}^m S_j)$-Lipschitz. We leave the possibility of extending our optimality relationships between the primal and multiradial dual to nonconvex optimization to future works. Doing so would likely rely on replacing concavity assumptions by strictly upper radiality as done in [10]. However, such nonconvex problems are beyond the scope of the algorithms and analysis considered herein. Lastly, note that x_{SL} will never be assumed to be known; it is only used in our analysis.

3.1 Exact MultiRadial Dual Optimality Relationships

These four assumptions suffice to show our primal and multiradially dual optimization problems are closely related. Our first result to this end is Theorem 3.1, which states that the two problems are equivalent when the optimal objective value is one, mirroring (3.3). This theorem is proved in Section 3.3.3.

Theorem 3.1. *Under assumptions A - D, if* $p^* = 1$ *or* $d^* = 1$, *then*

$$\operatorname{argmax} \Psi \times \{p^*\} = \operatorname{argmin} \Phi \times \{d^*\} \ .$$

Problems with any $p^* > 0$ (not necessarily one) are still amenable to the application of this result by considering the rescaled primal function Ψ_τ and its multiradially dual function Φ_τ, given by

$$\Psi_\tau(x) = \min \{\tau f(x), \hat{\iota}_{\mathcal{S}}(x)\} \tag{3.7}$$

$$\Phi_\tau(y) = \max\{(\tau f)^{\Gamma, e_0}(y), \varphi_{\mathcal{S}}(y)\} \tag{3.8}$$

for $\tau > 0$. We let $p(\tau)$ and $d(\tau)$ respectively denote

$$p(\tau) := \max_{x \in \mathcal{E}} \Psi_\tau(x) \tag{3.9}$$

$$d(\tau) := \min_{y \in \mathcal{E}} \Phi_\tau(y) \ . \tag{3.10}$$

By Theorem 3.1, $p(\tau) = d(\tau) = 1$ whenever $\tau = \frac{1}{p}$ and these problems have the same set of solutions. Since $\operatorname{argmax} \Psi = \operatorname{argmax} \Psi_\tau$, the following duality relation holds.

$$\operatorname{argmax} \Psi = \operatorname{argmin} \Phi_{1/p} \quad . \tag{3.11}$$

For algorithmic purposes, requiring knowledge of p^* is often prohibitive. As one example where such results are relevant, consider any minimization problem where strong duality holds. Then, minimizing the duality gap has a known optimal value, zero. To be concrete, consider a generic conic program over a closed convex cone $\mathcal{K}$ with dual cone $\mathcal{K}^*$ where the primal problem minimizes $\langle c, x \rangle$ subject to $Ax = b$ and $x \in \mathcal{K}$ and the dual problem maximizes $\langle b, y \rangle$ subject to $c - A^*y \in \mathcal{K}^*$. Then one can formulate seeking optimal primal-dual solutions as the following problem with $p^* = 1$

$$1 = \begin{cases} \max & 1 + \langle b, y \rangle - \langle c, x \rangle \\ \text{s.t.} & Ax = b \\ & x \in \mathcal{K} \\ & c - A^*y \in \mathcal{K}^* \ . \end{cases}$$

3.1.1 A Simple Method when the Optimal Value is Known When p^* is known and positive, (3.11) provides an alternative means to compute an approximate maximizer of the original problem. Given an initial point $x_0 \in \mathcal{E}$ and a given target accuracy $\varepsilon > 0$, one could iterate

$$\begin{cases} y_0 = \texttt{fom.initialize}(x_0, \varepsilon, \Phi_{1/p}\ ) \\ y_{i+1} = \texttt{fom.step}(y_i, \varepsilon, \Phi_{1/p}\ ) \end{cases} \quad . \tag{3.12}$$

Guarantees on this scheme's convergence directly follow from the convergence rate $K_{\texttt{fom}}(\cdot)$ of the given first-order method. The following theorem formalizes the primal objective gap and feasibility convergence of the above multiradial dual iterates y_i.

Theorem 3.2. *Under Assumptions A - D, the points* $z_i = e_0 + \frac{y_i - e_0}{\Phi_{1/p}\ (y_0)}$, *where* y_i *is the sequence* (3.12)*, have*

$$\frac{p^* - f(z_i)}{p^*} \leq \varepsilon \quad \textit{and} \quad \inf_{x \in S_0 \cap \mathcal{S}} \|z_i - x\| \leq \left[\frac{\varphi_{\mathcal{S}}(e_0) D_0}{1 - \varphi_{\mathcal{S}}(x_{SL})}\right] \varepsilon$$

for some $i \leq K_{\texttt{fom}}(\|x_0 - x^*\|, \varepsilon, \Phi_{1/p}\ )$.

Proof. Note some $i \leq K_{\texttt{fom}}(\|x_0 - x^*\|, \varepsilon, \Phi_{1/p}\ )$ must have $0 \leq \Phi_{1/p}\ (y_i) - 1 \leq \varepsilon$. The claimed objective bound on the corresponding z_i follows as

$$\frac{1}{p^*} f(z_i) \geq \limsup_{v \searrow \Phi_{1/p}\ (y_i)} \frac{1}{p^*} f\left(e_0 + \frac{y_i - e_0}{v}\right) \geq \frac{1}{\Phi_{1/p}\ (y_i)} = 1 - \frac{\Phi_{1/p}\ (y_i) - 1}{\Phi_{1/p}\ (y_i)} \geq 1 - \frac{\varepsilon}{\Phi_{1/p}\ (y_i)}$$

where first inequality uses upper semicontinuity, the second uses the definition of $(f/p^*)^{\Gamma, e_0}$, and the third uses that y_i is an ε-minimizer. The proof of our feasibility bound is deferred to Lemma 3.3 showing $\inf_{x \in S_0 \cap \mathcal{S}} \|z_i - x\| \leq \left[\frac{\varphi_{\mathcal{S}}(e_0) D_0}{1 - \varphi_{\mathcal{S}}(x_{SL})}\right] \frac{\Phi_{1/p}\ (y_i) - 1}{\Phi_{1/p}\ (y_i)}$. □

For example, consider the convex identifier $\varphi_S = \max\{\gamma_{S_j, e_j}\}$ as the maximum of the gauges of the constraint sets S_j with respect to e_j. Noting each gauge is $1/R_{e_j}(S_j)$-Lipschitz, the corresponding multiradial problem is $1/R$-Lipschitz where $R = \min_{j=0,\ldots,m} R_{e_j}(S_j)$. Consequently, a multiradial subgradient method (that is, using the subgradient method (2.17) in the multiradial method (3.12)) requires at most

$$K_{\texttt{subgrad}}(\|x_0 - x^*\|^2, \varepsilon, \Phi_{1/p}\ ) = \frac{\|x_0 - x^*\|^2}{R^2 \varepsilon^2}$$

iterations to produce some point with $\frac{p\ - f(z_i)}{p} \leq \varepsilon$ and $\inf_{x \in S_0 \cap \mathcal{S}} \|z_i - x\| \leq \left[\frac{\varphi_{\mathcal{S}}(e_0) D_0}{1 - \varphi_{\mathcal{S}}(x_{SL})}\right] \varepsilon$. Note this result is in line with prior radial subgradient method guarantees [5], avoiding reliance on Lipschitz constant assumptions and instead only depending on "geometric" radius and diameter-type constants. Unlike these prior methods, a common $e \in \text{int}\ \cap S_j$ is not needed and as previously noted, the value of R may be strictly larger.

3.2 General MultiRadial Dual Optimality Relationships

In the remainder of this section, we consider the relationship between p^* and d^* when p^* is unknown, so simply rescaling the objective to have optimal value one beforehand is not doable. Our Theorems 3.3 and 3.4 bound the absolute and relative distance of p^* and d^* from the value one in terms of each other. These theorems are proved in Section 3.3

Theorem 3.3. *Under Assumptions A - D, if* $p^* - 1 \geq 0$ *then*

$$1 - d^* \geq \frac{R_0 \eta}{R_0 + D_0} \frac{p^* - 1}{p^*}.$$

Theorem 3.4. *Under Assumptions A - D, if* $1 - d^* \geq 0$ *then*

$$p^* - 1 \geq \frac{R_0}{D_0 + R_0} \frac{1 - d^*}{d^*}.$$

In fact, if $y \in \mathcal{S}$ *satisfies* $f^{\Gamma, e_0}(y) \leq 1$, *then* $f(y) - 1 \geq \frac{R_0}{D_0 + R_0} \frac{1 - f^{\Gamma, e_0}(y)}{f^{\Gamma, e_0}(y)}$.

These two theorems provide bounds on the relative distance from the primal/dual optimal value from one in terms of the dual/primal's optimal value's absolute gap from one. Such conversions between absolute and relative accuracy have occurred throughout prior works on radial methods, see Renegar [3, 4]. For our multiradial theory, these relationships are primarily controlled by the natural geometric condition number based on the objective function's domain $R_0/(D_0 + R_0)$.

Consider applying these bounds to a rescaled problem with objective function τf for some $\tau \geq 1/p^*$. Recall this rescaled problem's maximum value is denoted by $p(\tau)$. In such rescaled settings, bounding $d^* \leq 1$, we denote the two coefficients above as

$$\rho = \frac{R_0}{D_0 + R_0} \quad \text{and} \quad c_\tau = \frac{1}{p(\tau)} \frac{R_0}{D_0 + R_0} \eta. \tag{3.13}$$

This notation helps illuminate the following relation implied by our theory

$$c_\tau [p(\tau) - 1] \leq 1 - d(\tau) \leq \frac{1}{\rho} [p(\tau) - 1] \quad \text{whenever } p(\tau) - 1 \geq 0 \text{ or } 1 - d(\tau) \geq 0. \tag{3.14}$$

The following corollaries of Theorems 3.3 and 3.4 provide the basis for our algorithms.

Corollary 3.1. *Let $\tau \geq 1/p^* > 0$ and $r \in [0, 1]$. Under Assumptions A - D, any y with $\Phi_\tau(y) - d(\tau) \leq (1 - r)[1 - d(\tau)]$ has $y \in \mathcal{S}$ and $\tau f(y) - 1 \geq r\rho[1 - d(\tau)]$.*

Proof. If $\Phi_\tau(y) - d(\tau) \leq (1 - r)[1 - d(\tau)]$, then $1 - \Phi_\tau(y) \geq r[1 - d(\tau)]$. Since $\tau \geq 1/p^*$ implies $1 - d(\tau) \geq 0$ by Theorem 3.3, it follows that $y \in \mathcal{S}$ as $\varphi_{\mathcal{S}}(y) \leq \Phi_\tau(y) \leq 1$. Moreover, by Theorem 3.4, $\tau f(y) - 1 \geq \rho[1 - \Phi_\tau(y)] \geq r\rho[1 - d(\tau)]$. □

Corollary 3.2. *Let $\tau_0 \geq 1/p^*$, $\delta \geq 0$, and $\mu \geq 1$. Under Assumptions A - D, if $\tau_1 \leq \frac{1}{1+\delta}\tau_0$ and $1 - d(\tau) \leq \mu\delta$ then*

$$p(\tau_1) - 1 \leq \frac{1}{1 + \delta} \left(1 - \frac{c_\tau}{\mu}\right) [p(\tau_0) - 1].$$

Proof. Applying first that $\tau_1 \leq \frac{1}{1+\delta}\tau_0$, second that $1 - d(\tau_0) \leq \mu\delta$, and third Theorem 3.3, one has

$$\begin{aligned} p(\tau_1) - 1 \leq \frac{1}{1 + \delta} p(\tau_0) - 1 = \frac{1}{1 + \delta} [p(\tau_0) - 1 - \delta] &\leq \frac{1}{1 + \delta} \left[p(\tau_0) - 1 - \frac{c_{\tau_0}}{\mu} \frac{1 - d(\tau)}{c_{\tau_0}}\right] \\ &\leq \frac{1}{1 + \delta} \left(1 - \frac{c_{\tau_0}}{\mu}\right) [p(\tau_0) - 1] \ . \end{aligned}$$

□

3.2.1 A Simple Method when Rescaled Problems can be Solved Exactly To demonstrate how to benefit algorithmically from Theorems 3.3 and 3.4, let $\tau_0 \geq 1/p^*$ and consider the sequence

$$\begin{cases} y^{(k+1)} \in \operatorname{argmin} \Phi_{\tau_k} \\ \tau_{k+1} = \frac{1}{f(y^{(k+1)})}. \end{cases} \tag{3.15}$$

With $r = 1$, Corollary 3.1 implies $\tau_{k+1} \leq \frac{1}{1+\rho[1-d(\tau_k)]}\tau_k$. Therefore, taking $\delta_k = \rho[1 - d(\tau_k)]$ and $\mu = 1/\rho$, we have $\tau_{k+1} \leq \frac{1}{1+\delta_k}\tau_k$ and $1 - d(\tau_k) \leq \mu\delta_k$ for all $k \geq 0$. This implies $p(\tau_{k+1}) - 1 \leq \frac{1}{1+\delta_k}(1 - \rho c_{\tau_k})[p(\tau_k) - 1]$ by Corollary 3.2. Noting $c_{\tau_k} \geq c_{\tau_0}$ since τ_k is decreasing and $\delta_k \geq 0$ yields the following theorem.

Theorem 3.5. *Under Assumptions A - D, if* $\tau_0 p^* - 1 > 0$*, the ideal sequence* (3.15) *has*

$$(1 - \rho c_{\tau_0})^k [\tau_0 p^* - 1] \geq p(\tau_k) - 1.$$

Hence for any $\varepsilon > 0$*, all* $k \geq \frac{1}{\rho c_{\tau_0}} \log(p^*[\tau_0 p^* - 1]/\varepsilon)$ *have*

$$p^* - f(y^{(k)}) < \varepsilon \quad \textit{and} \quad y^{(k)} \textit{ feasible.}$$

3.2.2 A Simple Method when Rescaled Problems are Solved Inexactly The sequence (3.15) will often not be practical to implement as it requires exact solutions to the multiradial dual problems $\min_{y \in \mathcal{E}} \Phi_\tau(y)$. However, the linear convergence in Theorem 3.5 suggests that a good primal solution may be obtained by approximately solving a (relatively) small number of dual problems. This is the main motivation behind our general multiradial methods; we mimic the sequence (3.15), replacing exact solutions with approximate ones.

Here we sketch a general family of methods of this form, with the drawback that determining when an approximate solution is good enough still requires unrealistic problem-dependent knowledge. We suppose an initial feasible point $y^{(0)} \in \text{dom } f \cap \mathcal{S}$ is given and set $\tau_0 = 1/f(y^{(0)})$. To approximate the iteration (3.15), we apply a given `fom` to minimize Φ_{τ_k} initialized at $y^{(k)}$, yielding iterates $y_i^{(k)}$. Once a sufficient accuracy is reached at some iteration i of the subproblem optimization, we set $y^{(k+1)} = y_i^{(k)}$. One natural way to define sufficient accuracy is to require $\Phi_{\tau_k}(y_i^{(k)}) - d(\tau_k) \leq \delta_k$. If $d(\tau_k) + \delta_k \leq 1$, then such $y_i^{(k)}$ will be feasible and Theorem 3.4 implies $1 + \rho\delta_k \leq \tau_k f(y_i^{(k)})$. Motivated by this observation, we bypass the 'dual' notion of accuracy and directly say $y_i^{(k)}$ is sufficiently accurate if (i) $1/f(y_i^{(k)}) \leq \frac{1}{1+\delta_k}\tau_k$ and (ii) $y_i^{(k)} \in \cap_{j=1}^m S_j$. This process is formalized in Algorithm 1.

Algorithm 1 The MultiRadial Method

Require: (f, e_0), $x_0 \in \text{dom } f \cap \mathcal{S}$, a convex identifier $\varphi_{\mathcal{S}}$, $\{\delta_k\}_{k=0}^{\infty}$, a first-order method `fom`

Set $\tau_0 = 1/f(x_0)$ and $y_0^{(0)} = \texttt{fom.initialize}(x_0, \delta_0, \Phi_{\tau_0})$ and $i = 0$
for $k = 0, 1, 2, \ldots,$ **do**
  repeat
    $y_{i+1}^{(k)} = \texttt{fom.step}(y_i^{(k)}, \delta_k, \Phi_{\tau_k})$, $i = i + 1$ (`fom` takes one step)
  until $1/f(y_i^{(k)}) \leq \frac{1}{1+\delta_k}\tau_k$ and $y_i^{(k)}$ is feasible
  Set $\tau_{k+1} = 1/f(y_i^{(k)})$ (Restart `fom` once satisfied by $y_i^{(k)}$)
  Set $y_0^{(k+1)} = \texttt{fom.initialize}(y_i^{(k)}, \delta_{k+1}, \Phi_{\tau_{k+1}})$ and $i = 0$
end for

Selecting a sequence of stopping criteria δ_k for which Algorithm 1 has provably good performance guarantees is nontrivial. As $1 - d(\tau_k)$ decreases to 0, δ_k must decrease similarly for the condition $1/f(y_i^{(k)}) \leq \frac{1}{1+\delta_k}\tau_k$ to be attainable; below we show that taking $\delta_k = \rho\frac{1-d(\tau_k)}{2}$ maintains the outer linear convergence rate of (3.15).

Theorem 3.6. *Suppose Assumptions A - D hold and let* `fom` *be given. Then, for all* $\varepsilon > 0$*, Algorithm 1 with* $\delta_k = \rho\frac{1-d(\tau_k)}{2}$ *has*

$$p^* - f(y_0^{(k')}) \leq \varepsilon \quad \textit{and} \quad y_0^{(k')} \textit{ feasible}$$

for some $k' \leq N := \lceil \frac{2}{\rho c_{\tau_0}} \log(\frac{p^*[\tau_0 p^* - 1]}{\varepsilon}) \rceil$*. The total number of* `fom` *steps needed to find such* $y_0^{(k')}$ *is at most* $\sum_{k=1}^{N} K_{\texttt{fom}}(D, \delta_k/\rho, \Phi_{\tau_k})$.

Proof. Our proof is split into two parts. First, we bound the number of inner loop steps before the stopping criterion is met using Corollary 3.1. Then, we bound the number of outer loop steps before an ε-minimizer by a similar contraction as seen for the exact method (3.15) using Corollary 3.2.

By definition, the first-order method `fom` must have some iteration i_k attain $\Phi_{\tau_k}(y_{i_k}^{(k)}) - d(\tau_k) \leq \delta_k/\rho$ with $i_k \leq K_{\texttt{fom}}(\|y_0^{(k)} - y_*^{(k)}\|, \delta_k/\rho, \Phi_{\tau_k}) \leq K_{\texttt{fom}}(D, \delta_k/\rho, \Phi_{\tau_k})$. Since $\delta_k/\rho = \frac{1-d(\tau_k)}{2}$, Corollary 3.1 implies $y_{i_k}^{(k)}$ is feasible and $\tau_k f(y_{i_k}^{(k)}) - 1 \geq (1 - 1/2)\rho(1 - d(\tau_k)) = \delta_k$ (or equivalently, $1/f(y_{i_k}^{(k)}) \leq \frac{\tau_k}{1+\delta_k}$). Therefore, $y_{i_k}^{(k)}$ would satisfy the stopping criteria for the inner loop of Algorithm 1 and so, the inner loop at iteration k will always terminate within $K_{\texttt{fom}}(D, \delta_k/\rho, \Phi_{\tau_k})$ steps.

Notice that the inner loop stopping criterion ensures that $\tau_{k+1} \leq \frac{1}{1+\delta_k}\tau_k$, where $1 - d(\tau_k) \leq \frac{2}{\rho}\delta_k$. Therefore, each outer loop contracts the (rescaled) objective gap towards one since

$$\begin{aligned} p(\tau_{k+1}) - 1 \leq \frac{1}{1+\delta_k}\left(1 - \frac{\rho c_{\tau_k}}{2}\right)[p(\tau_k) - 1] &\leq \frac{1}{1+\delta_k}\left(1 - \frac{\rho c_{\tau_0}}{2}\right)[p(\tau_k) - 1] \\ &\leq \left(1 - \frac{\rho c_{\tau_0}}{2}\right)[p(\tau_k) - 1], \end{aligned}$$

where the first inequality used Corollary 3.2 and the second used that c_{τ_k} is increasing. As such, the primal gap converges linearly with $p(\tau_k) - 1 \leq (1 - \frac{\rho c_{\tau_0}}{2})^k[\tau_0 p^* - 1]$ and consequently, some $k' \leq \lceil \frac{2}{\rho c_{\tau_0}} \log(\frac{p\ [\tau_0 p\ -1]}{\varepsilon}) \rceil$ has a feasible $y_0^{(k\ )}$ with $p^* - f(y_0^{(k\ )}) \leq \varepsilon$. Totalling the number of steps executed by `fom` to find each $y_0^{(k)}$ gives the claim. □

3.3 Proofs for MultiRadial Duality Theory Optimality Relationships

Below, we first prove Theorems 3.3 and 3.4, which bound the primal and dual difference from one in terms of each other. From these, Theorem 3.1 is almost immediate.

Our proofs use the following two facts repeatedly. Under Assumptions A - D,

$$f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y) \qquad \forall x, y \in S_0, \tag{3.16}$$

$$f(e_0 + \frac{y - e_0}{v}) \geq \frac{1}{v} \qquad \text{for all positive } v \geq f^{\Gamma, e_0}(y). \tag{3.17}$$

3.3.1 Proof of Theorem 3.3 This result primarily follows from the following bound on the radial dual value at x^*

$$f^{\Gamma,e_0}(x^*) \leq 1 - (1 - \gamma_{S_0,e_0}(x^*))\frac{p^* - 1}{p^*}. \tag{3.18}$$

We delay the proof of this inequality to first show it suffices to prove the theorem. Consider $x_\lambda = \lambda x_{SL} + (1-\lambda)x^*$ with $\lambda = \frac{R_0}{R_0 + D_0}(1 - \gamma_{S_0,e_0}(x^*))\frac{p\ -1}{p}$. This satisfies

$$f^{\Gamma,e_0}(x_\lambda) \leq f^{\Gamma,e_0}(x^*) + \lambda\frac{D_0}{R_0} \leq 1 - \frac{R_0\eta}{R_0 + D_0}\frac{p^* - 1}{p^*}$$

where the first inequality uses the $1/R_0$-Lipschitz continuity of f^{Γ,e_0} and that $\|x_{SL} - x^*\| \leq D_0$, and the second uses (3.18) and that $(1 - \gamma_{S_0,e_0}(x^*)) \geq \eta$. From the convexity of $\varphi_{\mathcal{S}}$, it follows that

$$\varphi_{\mathcal{S}}(x_\lambda) \leq 1 - \frac{R_0\eta}{R_0 + D_0}\frac{p^* - 1}{p^*}.$$

Combined, these two bounds prove the result as $\Phi(x_\lambda) \leq 1 - \frac{R_0\eta}{R_0+D_0}\frac{p\ -1}{p}$.

Now we return to prove (3.18). Consider $\hat{v} \in (\gamma_{S_0,e_0}(x^*), 1)$ and let $\hat{x} = e_0 + \frac{x\ -e_0}{\hat{v}} \in S_0$. For any $v \in [\hat{v}, 1]$, note that $e_0 + \frac{x\ -e_0}{v} = (1-\alpha)\hat{x} + \alpha x^*$ where $\alpha = 1 - \frac{\hat{v}(1-v)}{v(1-\hat{v})} \in [0,1]$. Since $x^*, \hat{x} \in S_0$, it follows that

$$f\left(e_0 + \frac{x^* - e_0}{v}\right) \geq \alpha p^* + (1-\alpha) f(\hat{x}) \geq \alpha p^* = \frac{v - \hat{v}}{v(1-\hat{v})} p^*$$

where the first inequality uses (3.16) and the second uses $f(\hat{x}) \geq 0$. In particular, for $v \in [\hat{v} + [1 - \hat{v}]/p^*, 1]$, this ensures $vf\left(e_0 + \frac{x\ -e_0}{v}\right) \geq 1$. Hence $f^{\Gamma,e_0}(x^*) \leq 1 - (1-\hat{v})(p^*-1)/p^*$ since f is strictly upper radial. Taking the limit as $\hat{v} \to \gamma_{S_0,e_0}(x^*)$ gives the claim.

3.3.2 Proof of Theorem 3.4 We separate the proof into two lemmas which are of interest in their own right. The first lemma shows that the multiradial dual function has bounded level sets. Since this function is convex and bounded below, the lemma guarantees that the function has global minimizers. The second lemma establishes the inequality $f(y) - 1 \geq \frac{R_0}{R_0+D_0}\frac{1-f^{\Gamma,e_0}(y)}{f^{\Gamma,e_0}(y)}$. Combining these two lemmas immediately completes the proof.

Lemma 3.3. *Under Assumptions A - D, if $\Phi(y) \leq 1+\varepsilon$ for $\varepsilon \geq 0$, then $y_\varepsilon = e_0 + \frac{y-e_0}{1+\varepsilon}$ has $\inf_{x \in S_0 \cap \mathcal{S}} \|y_\varepsilon - x\| \leq \frac{\varepsilon}{1+\varepsilon}\frac{\varphi_S(e_0)}{1-\varphi_S(x_{SL})} D_0$. In particular, $\|y - e_0\| \leq D_0 + \varepsilon\left[1 + \frac{\varphi_S(e_0)}{1-\varphi_S(x_{SL})}\right] D_0$.*

Proof of Lemma 3.3. Consider the point $x_\lambda = (1-\lambda)x_{SL} + \lambda y_\varepsilon$ with $\lambda = \frac{(1-\varphi_S(x_{SL}))(1+\varepsilon)}{(1-\varphi_S(x_{SL}))(1+\varepsilon)+\varepsilon\varphi_S(e_0)} \in [0,1]$. First, observe $x_\lambda \in S_0$ follows from convexity of S_0 since $x_{SL} \in S_0$ by definition and $y_\varepsilon \in S_0$ by (3.17) noting $f^{\Gamma,e_0}(y) \leq \Phi(y) \leq 1+\varepsilon$. Next, observe $x_\lambda \in \mathcal{S}$ by our choice of λ as

$$\begin{aligned}\varphi_{\mathcal{S}}(x_\lambda) &\leq \varphi_{\mathcal{S}}(x_{SL}) + \lambda[\varphi_{\mathcal{S}}(y_\varepsilon) - \varphi_{\mathcal{S}}(x_{SL})] \\ &\leq \varphi_{\mathcal{S}}(x_{SL}) + \lambda\left[1 + \frac{\varepsilon}{1+\varepsilon}\varphi_{\mathcal{S}}(e_0) - \varphi_{\mathcal{S}}(x_{SL})\right] = 1\end{aligned}$$

where the first inequality uses convexity of $\varphi_{\mathcal{S}}$ at x_λ, and the second uses convexity of $\varphi_{\mathcal{S}}$ at y_ε and that $\varphi_{\mathcal{S}}(y) \leq \Phi(y) \leq 1+\epsilon$. Together since $x_\lambda \in S_0 \cap \mathcal{S}$, we conclude

$$\inf_{x \in S_0 \cap \mathcal{S}} \|y_\varepsilon - x\| \leq \|y_\varepsilon - x_\lambda\| = \frac{\varepsilon}{1+\varepsilon}\frac{\varphi_{\mathcal{S}}(e_0)}{1-\varphi_{\mathcal{S}}(x_{SL})}\|x_\lambda - x_{SL}\|.$$

Bounding $\|x_\lambda - x_{SL}\| \leq D_0$ gives the lemma's first claim. The second claim follows similarly, noting $\|x_\lambda - e_0\| \leq D_0$ as well and applying the triangle inequality

$$\begin{aligned}\|y - e_0\| = (1+\varepsilon)\|y_\varepsilon - e_0\| &\leq (1+\varepsilon)(\|y_\varepsilon - x_\lambda\| + \|x_\lambda - e_0\|) \\ &\leq D_0 + \varepsilon\left[1 + \frac{\varphi_{\mathcal{S}}(e_0)}{1-\varphi_{\mathcal{S}}(x_{SL})}\right] D_0.\end{aligned}$$

□

Lemma 3.4. *Suppose hypo f is convex and f is globally upper semi-continuous. Then, for any $e_0 \in$ dom f, $0 \leq f^{\Gamma,e_0}(y) \leq 1$ implies $f(y) - 1 \geq \frac{(1-f^{\Gamma,e_0}(y))R_0}{\|y-e_0\|+f^{\Gamma,e_0}(y)R_0}$.* [4] *In addition, if $D_0 := D(S_0) < \infty$, then $f(y) \geq 1 + \frac{R_0}{D_0+R_0}\frac{1-f^{\Gamma,e_0}(y)}{f^{\Gamma,e_0}(y)}$.*

[4] If $R_0 = \infty$ this should be taken as $f(y) \geq 1 + \frac{1-f^{\Gamma,e_0}(y)}{f^{\Gamma,e_0}(y)}$

Proof of Lemma 3.4. Fix y with $0 \leq 1 - f^{\Gamma,e_0}(y) \leq 1$. If $R_0 = \infty$ then f is strictly positive and concave on $\mathcal{E}$ which is possible only if f is constant. If f is constant, then so is f^{Γ,e_0}, hence $f^{\Gamma,e_0}(y) = f^{\Gamma,e_0}(e_0) = \frac{1}{f(e_0)}$. Therefore, if $R_0 = \infty$ or $y = e_0$, then $f(y) = 1/f^{\Gamma,e_0}(y) = 1 + \frac{1-f^{\Gamma,e_0}(y)}{f^{\Gamma,e_0}(y)}$.

Suppose for the rest of the proof that $y \neq e_0$ and $R_0 < \infty$. Let $z = e_0 - R_0 \frac{y-e_0}{\|y-e_0\|} \in S_0$. For any positive $v \geq f^{\Gamma,e_0}(y)$, let $w = e_0 + \frac{y-e_0}{v}$ and $\lambda = v\left(1 + \frac{R_0(1-v)}{\|y-e_0\|+R_0 v}\right)$. Then noting $y = \lambda w + (1-\lambda)z$, it follows that

$$f(y) = f(\lambda w + (1-\lambda)z) \geq \lambda f(w) + (1-\lambda)f(z) \geq \frac{\lambda}{v} = 1 + \frac{R_0(1-v)}{\|y-e_0\| + R_0 v}$$

where the first inequality uses (3.16), and the second uses (3.17) and that $f(z) \geq 0$. Taking the limit as $v \to f^{\Gamma,e_0}(y)$ gives $f(y) \geq \frac{R_0(1-f^{\Gamma,e_0}(y))}{\|y-e_0\|+R_0 f^{\Gamma,e_0}(y)}$. The second part of the lemma follows from

$$f^{\Gamma,e_0}(y) \geq \gamma_{S_0,e_0}(y) \geq \frac{1}{D_0}\|y - e_0\|,$$

where the first inequality follows from (3.17) and the second follows because if $S_0 \subset B(e_0, D_0) := \{x \in \mathcal{E} \mid \|x - e_0\| \leq D_0\}$, then $\frac{1}{D_0}\|\cdot - e_0\| = \gamma_{B(e_0,D_0),e_0} \leq \gamma_{S_0,e_0}$. □

3.3.3 Proof of Theorem 3.1 Theorems 3.3 and 3.4 imply that $p^* = 1$ if and only if $d^* = 1$. It remains to show that the two problems have the same solutions. To that end, suppose $x^* \in \mathcal{S}$ and $f(x^*) = p^* = 1$. Since f is strictly upper radial, $f(x^*) = 1$ implies $f^{\Gamma,e_0}(x^*) = 1$. From $x^* \in \mathcal{S}$, we get $\varphi_{\mathcal{S}}(x^*) \leq 1$, hence $\Phi(x^*) = f^{\Gamma,e_0}(x^*) = 1 = d^*$. On the other hand, suppose $\Phi(y^*) = d^* = 1$. Then, $\varphi_{\mathcal{S}}(y^*) \leq 1$ hence $y^* \in \mathcal{S}$ and $f(y^*) \leq p^* = 1$. We also have $f^{\Gamma,e_0}(y^*) \leq 1$ which, by the second part of Theorem 3.4, implies $f(y^*) \geq 1 + \frac{R_0}{R_0+D_0}(1 - f^{\Gamma,e_0}(y^*)) \geq 1$. Therefore, $f(y^*) = 1 = p^*$, and the proof is complete.

4 A Parameter-Free, Optimal, Parallel MultiRadial Method

The previously discussed multiradial method in Algorithm 1 required unrealistic knowledge to compute the needed δ_k. In this section, we present a parameter-free adaption of this method, using the parallel restarting ideas of [24], which we call the Parallel MultiRadial Method (||-MRM).

Conceptually, the ||-MRM can be thought of as consisting of N parallel, but not independent, instances of Algorithm 1. In each l-th instance, Algorithm 1 is run with a constant accuracy sequence $\left\{\delta_k^{(l)}\right\}_{k=0}^{\infty} = \left\{\delta^{(l)}\right\}$. All instances start with the same initial data (f, e_0), convex identifier $\varphi_{\mathcal{S}}$, feasible x_0, and fom. With a slight abuse of notation, we denote each instance by $\texttt{fom}^{(l)}$. The crucial part of ||-MRM is that the instances cooperate by sharing their feasible iterates with one another. In particular, one instance, say $\texttt{fom}^{(1)}$, can use an iterate of another, say $\texttt{fom}^{(2)}$, to make the update in step 6 of Algorithm 1, if such an iterate is sufficiently accurate for $\texttt{fom}^{(1)}$. In the end, the best feasible iterate among all is returned as the solution. The number of instances N and the respective target accuracy $\delta^{(l)}$ for each instance can be treated as inputs to the method. A concrete implementation of the ||-MRM is given in Algorithm 2. For simplicity, Algorithm 2 takes $b \geq 2$ and N as inputs and automatically sets $\delta^{(l)} = b^{-l}$. Motivated by Theorem 3.6, we find that setting $N = O(\log_b(1/\varepsilon))$ is sufficient to reach ε-accuracy.

More formally, Algorithm 2 produces iterates $y_i^{(l)}$, $i = 0, 1, \ldots$, $l = 1, \ldots, N$. For each l, the next iterate $y_{i+1}^{(l)}$ is produced from the previous one by a single step of $\texttt{fom}^{(l)} = \texttt{fom}$, i.e., $y_{i+1}^{(l)} = \texttt{fom}^{(l)}.\texttt{step}(y_i^{(l)}, \delta^{(l)}, \Phi_{\tau_i^{(l)}})$. These steps can be done in parallel or, as described in Algorithm 2,

sequentially. Then the best iterate among all past and present feasible iterates, denoted y_{i+1}^{best}, is computed. With $\tau_{i+1}^{best} = 1/f(y_{i+1}^{best})$, each instance $\texttt{fom}^{(l)}$ for which y_{i+1}^{best} meets their restarting criteria (e.g., $\tau_{i+1}^{best} \leq \frac{1}{1+\delta^{(l)}}\tau_i^{(l)}$), will set their $\tau_{i+1}^{(l)}$ as τ_{i+1}^{best} and reinitialize $\texttt{fom}^{(l)}$ at y_{i+1}^{best}. We refer to this event as $\texttt{fom}^{(l)}$ restarting.

Algorithm 2 The Parallel MultiRadial Method (||-MRM)

Require: (f, e_0), $x_0 \in \operatorname{dom} f \cap \mathcal{S}$, a convex identifier $\varphi_{\mathcal{S}}$, $b \geq 2$, $N > 0$, a first-order method fom

Set $\delta^{(l)} = b^{-l}$ for each $l = 1, \ldots, N$
Set each $\tau_0^{(l)} = 1/f(x_0)$ and $y_0^{(l)} = \texttt{fom}^{(l)}.\texttt{initialize}(x_0, \delta_0^{(l)}, \Phi_{\tau_0^{(l)}})$
for $i = 0, 1, 2, \ldots,$ **do**
  for $l = 1 \ldots N$ **do**
    $\tau_{i+1}^{(l)} = \tau_i^{(l)}$ (Each $\texttt{fom}^{(l)}$ takes one step)
    $y_{i+1}^{(l)} = \texttt{fom}^{(l)}.\texttt{step}(y_i^{(l)}, \delta^{(l)}, \Phi_{\tau_i^{(l)}})$
  end for
  $y_{i+1}^{best} = \operatorname{argmin}\{1/f(y_i^{(l)}) \mid y_i^{(l)}$ is feasible and $i' \leq i+1\}$ (Find the best iterate seen thus far)
  $\tau_{i+1}^{best} = 1/f(y_{i+1}^{best})$
  for each $l = 1 \ldots N$ with $\tau_{i+1}^{best} \leq \frac{1}{1+\delta^{(l)}}\tau_i^{(l)}$ **do**
    Set $\tau_{i+1}^{(l)} = \tau_{i+1}^{best}$ (Restart $\texttt{fom}^{(l)}$ if satisfied by y_{i+1}^{best})
    Set $y_{i+1}^{(l)} = \texttt{fom}^{(l)}.\texttt{initialize}(y_{i+1}^{best}, \delta^{(l)}, \Phi_{\tau_{i+1}^{(l)}})$
  end for
end for

4.1 Convergence Guarantees and Theory

For ease of exposition, we define a few additional quantities not explicitly used in ||-MRM: For each instance of the first-order method l, we let $i_0^{(l)} < i_1^{(l)} < i_2^{(l)} < \cdots < i_{K_l}^{(l)}$ be the sequence of iterations where a restart occurred (i.e., $\tau_{i_k^{(l)}+1}^{best} \leq \frac{1}{1+\delta^{(l)}}\tau_{i_k^{(l)}}^{(l)}$). Note there must only be a finite number of such events. Each restart has

$$\tau_{i_k^{(l)}+1}^{(l)} \leq \frac{1}{1+\delta^{(l)}}\tau_{i_k^{(l)}}^{(l)} .$$

Inductively applying this and noting $1/f(y_{i+1}^{best}) \geq 1/p^*$, the total number of restarts by first-order method instance l is at most $K_l \leq \log(p^*/f(x_0))/\log(1+\delta^{(l)})$. For each iteration i, we say the *critical first-order method instance* l_i is the one with $\frac{b\delta^{(l_i)}}{\rho} \leq 1 - d(\tau_i^{best}) < \frac{b^2\delta^{(l_i)}}{\rho}$.

The following three quantities are useful in formalizing our convergence rate guarantees for Algorithm 2. They correspond to bounds on how long it takes for the instance l to be guaranteed to reach a $\delta^{(l)}$-minimizer of any of its subproblem, the first parallel instance that could be critical, and given some $\varepsilon > 0$, the last parallel instance that can be critical before an ε-minimizer is found.

$$K_{\texttt{fom}}^{(l)} := \max_{i \geq 0} K_{\texttt{fom}}(D_0, \delta^{(l)}/\rho, \Phi_{\tau_i^{(l)}}) \tag{4.1}$$

$$l_0 := \min\left\{ l = 1, 2, \cdots \mid \frac{b\delta^{(l)}}{\rho} < 1 \right\} = \lfloor \log_b(b^2/\rho) \rfloor \tag{4.2}$$

$$\tilde{N}(\varepsilon) := \max\left\{ l = 1, 2, \cdots \mid \frac{b^2\delta^{(l)}}{\rho} \geq \frac{c_{\tau_0}\varepsilon}{p^*} \right\} = \left\lfloor \log_b\left(\frac{b^2 p^*}{c_{\tau_0}\rho\varepsilon} \right) \right\rfloor . \tag{4.3}$$

Based on these, we have the following convergence guarantee (proof deferred to Section 6), establishing that ||-MRM needs a logarithmic number $O(\tilde{N}(\varepsilon))$ of multiradial dual problem solves, each requiring a number of steps controlled by the chosen first-order method, $K^{(l)}_{\texttt{fom}}$.

Theorem 4.1. *Suppose Assumptions A - D hold and let* fom *be given. Then for all* $\varepsilon > 0$, *Algorithm 2 with* fom, $b \geq 2$, *and* $N \geq \log_b(\frac{bp}{c_{\tau_0}\rho\varepsilon})$ *has*

$$p^* - f(y_i^{best}) \leq \varepsilon \quad \textit{and} \quad y_i^{best} \textit{ feasible}$$

provided that[5]

$$i \geq \frac{\log(\tau_0 p^*)}{\log(1+\frac{\rho}{b^2})} K^{(l_0)}_{\texttt{fom}} + \frac{5b^2(1+\frac{\rho}{b^2})}{4\rho c_{\tau_0}} \sum_{l=l_0+1}^{\tilde{N}(\varepsilon)} K^{(l)}_{\texttt{fom}} \ .$$

To illustrate the reach of this theorem, we present corollaries for three pairs of first-order methods (previously introduced in Section 2) and appropriately chosen convex identifiers. Detailed proofs of these corollaries are deferred to Appendix A. Suppose points $e_j \in \operatorname{int} S_j$ are known. First, noting that the gauges γ_{S_j,e_j} are uniformly Lipschitz, one may reasonably consider

$$\texttt{fom} = \texttt{subgrad} \qquad \text{and} \qquad \varphi_{\mathcal{S}} = \max\{\gamma_{S_1,e_1}, \ldots, \gamma_{S_m,e_m}\} \ . \tag{4.4}$$

A projected subgradient method requires $O(1/\varepsilon^2)$ subgradient evaluations to minimize a generic Lipschitz function. Applying Theorem 4.1, we find a parallel multiradial subgradient method also only requires $O(1/\varepsilon^2)$ subgradient evaluations, up to a parallelizable factor of $N = O(\log(1/\varepsilon))$.

Corollary 4.1. *Let Assumptions A - D hold. If* fom *and* $\varphi_{\mathcal{S}}$ *are set as in* (4.4)*, then, for all* $\varepsilon > 0$, *Algorithm 2 with* $N \geq \log_b(\frac{bp}{c_{\tau_0}\rho\varepsilon})$ *finds a feasible* ε*-minimizer within* $O(1/\varepsilon^2)$ *iterations.*

Much like prior radial methods [3, 5], no Lipschitz continuity assumptions are needed and projections are entirely avoided, instead just relying on linesearches and normal vectors. Unlike these prior radial methods, the usage of a common reference point $e \in \operatorname{int} \mathcal{S}$ is avoided. Instead, separate centers are utilized, which also facilitates the potential for smaller Lipschitz constants for each gauge.

If, additionally, the objective f is smooth and the sets S_j are smooth and compact, accelerated methods can be applied. A set is β-smooth if its unit normal vectors are β-Lipschitz continuous on the set's boundary. Recently, [9, Corollary 3.2] showed every β-smooth compact set S_j has $\frac{1}{2}\gamma^2_{S_j,e_j}$ as a $O(\beta)$-smooth function. Using this line of reasoning, one can also show that f^{Γ,e_0} is $O(\beta)$-smooth if f is β-smooth and dom f is compact. This motivates the usage of accelerated smoothing and generalized gradient methods applied with gauges squared occurring in the identifier. We consider the following two settings:

$$\begin{aligned} \texttt{fom} = \texttt{smooth} \qquad & \text{and} \qquad \varphi_{\mathcal{S}} = \max\{\varphi_{S_1}, \ldots, \varphi_{S_m}\} \\ & \text{where} \quad \varphi_{S_j}(x) = \begin{cases} \gamma_{S_j,e_j}(x) & \text{if } \gamma_{S,e}(x) > 1 \\ \frac{1}{2}\gamma^2_{S_j,e_j}(x) + \frac{1}{2} & \text{otherwise} \end{cases}, \end{aligned} \tag{4.5}$$

$$\texttt{fom} = \texttt{genGrad} \qquad \text{and} \qquad \varphi_{\mathcal{S}} = \max\left\{\gamma^2_{S_1,e_1}, \ldots, \gamma^2_{S_m,e_m}\right\} \ . \tag{4.6}$$

In both cases, up to a logarithmic term, these methods' $O(1/\varepsilon)$ and $O(1/\sqrt{\varepsilon})$ convergence rates are preserved, now providing new accelerated projection-free methods. For ease, our corollaries assume f is twice continuously differentiable. However, this assumption is not needed because, as pointed out earlier, f^{Γ,e_0} is smooth as long as f is smooth and dom f is compact.

[5] We use the convention that $\sum_{l=a}^{b} K^{(l)}_{\texttt{fom}} = 0$ if $b < a$.

Corollary 4.2. *Let Assumptions A - D hold and f be twice continuously di erentiable. If f is β-smooth and each S_j is β-smooth and compact, and* `fom` *and $\varphi_{\mathcal{S}}$ are set as in* (4.5)*, then, for all $\varepsilon > 0$, Algorithm 2 with $N \geq \log_b(\frac{bp}{c\tau_0\rho\varepsilon})$ finds a feasible ε-minimizer within $O(1/\varepsilon)$ iterations.*

Corollary 4.3. *Let Assumptions A - D hold and f be twice continuously di erentiable. If f is β-smooth and each S_j is β-smooth and compact, and* `fom` *and $\varphi_{\mathcal{S}}$ are set as in* (4.6)*, then, for all $\varepsilon > 0$, Algorithm 2 with $N \geq \log_b(\frac{bp}{c\tau_0\rho\varepsilon})$ finds a feasible ε-minimizer within $O(1/\sqrt{\varepsilon})$ iterations.*

4.2 Practical Consideration

To apply ||-`MRM` with $\varphi_{\mathcal{S}}$ constructed from gauges (squared) requires three main ingredients, computing the reference points e_j each interior to the related constraint S_j, computing a feasible initialization x_0 and $\tau_0 = 1/f(x_0) > 0$, and computing function values and subgradients of f^{Γ,e_0} and γ_{S_j,e_j} for the underlying first-order method. Below, we address these three computations and provide an extension to allow affine constraints (which have no interior and so are beyond the scope of Assumption A).

Computing Selection of e_j Our multiradial duality theory avoids a reliance on knowing a good reference point interior to $\cap S_j$ (with the quality measured by $R_e(\cap S_j)$). Instead, points e_j with reasonably positive $R_{e_j}(S_j)$ are needed. One natural choice of e_j is the Chebyshev center, defined as maximizing $e \mapsto R_e(S_j)$. For generic convex S_j, computing this is a convex optimization problem. For polyhedrons, this corresponds to an LP. For norm-type constraints $\{x \mid \|A_j x - b_j\| \leq 1\}$, its center is given by any solution to $A_j x = b_j$.

For our numerics, we consider QCQPs where the center is also given by a linear system solve. Our results in Section 5 show that in this setting, the choice of centers has no observable effect on convergence. Therefore, exact solutions to the systems $A_j x = b_j$ are not needed. One could, for instance, compute the centers e_j using only a few conjugate gradient steps. Note for a given e_j, computing or estimating $R_{e_j}(S_j)$ is nontrivial. For convex QCQPs (1.4), this amounts to a nonconvex QCQP.

Computing an initialization and rescaling τ_0 Given a selection e_j, ||-`MRM` still requires knowledge of a sufficiently large τ_0 such that $p(\tau_0) \geq 1$. This can be done directly by finding any $x_0 \in \cap_{j=0}^m S_j$ and setting $\tau_0 = 1/f(x_0)$. Such a point can be found by minimizing the maximum of the gauges of each S_j with respect to e_j until a value less than one is reached (which the Slater point ensures is possible). Noting that $\lim_{\tau\to\infty}(\tau f)^{\Gamma,e_0}(y) = \gamma_{S_0,e_0}(y)$, computing an initial feasible point in the domain of f can be viewed as approximately minimizing $\Phi_\infty(y) := \lim_{\tau\to\infty} \Phi_\tau(y)$. Hence the cost of adding such a first phase to bootstrap ||-`MRM` is comparable to the cost of approximately minimizing one subproblem Φ_τ.

Computing f^{Γ,e_0} and γ_{S_j,e_j} (and their subgradients) Often f^Γ and γ_S have closed forms (see [6, Tables 1 and 2]) and their (sub)gradients can be directly computed from (sup)gradients and normal vectors of f and S (see [10, Proposition 19 and 21]). For example, generic polyhedral constraints $\{x \mid Ax \leq b\}$ or ellipsoidal constraints $\{x \mid \|Ax - b\|_2 \leq 1\}$ have closed forms for their gauge, computable by a single matrix-vector multiplication, see (1.6).

If a closed form is not available, evaluating the radial transformation of a function or the gauge of a set amounts to a one-dimensional linesearch. Given a function value oracle for f or membership oracle for S_j, this can be computed by any root-finding methods (e.g., bisection). Algorithms based on such inexact evaluations were developed by the works [7, 8].

Reformulations with Affine Constraints As stated, our multiradial duality theory does not directly apply to problems with affine constraints $Ax = b$ among the set constraints $x \in \cap_{j=1}^m S_j$ since the affine constraints have no interior (and hence cannot satisfy Assumption A). Such constraints can be addressed separately from S_j by additionally requiring that each e_j satisfies $Ae_j = b$, then consider the affine constrained primal and multiradial dual functions

$$\Psi(x) = \begin{cases} f(x) & \text{if } Ax = b,\ x \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases} \qquad \Phi(y) = \begin{cases} \max\{f^{\Gamma,e_0}(x), \gamma_{S_j,e_j}(y)\} & \text{if } Ay = b \\ \infty & \text{otherwise .} \end{cases}$$

Our Theorems 3.1, 3.3, and 3.4 directly generalize to this setting which restricts to the affine subspace where $Ax = b$, relating maximizers of Ψ and minimizers of Φ. Consequently, given a first-order method capable of minimizing a finite maximum over affine constraints, ||-`MRM` could be applied to solve an affine-constrained primal. For example, by precomputing the projection operator onto the affine space, a projected subgradient method could be applied, while remaining projection-free with respect to the more sophisticated S_j constraints.

5 Numerical Validation

In this final section, we apply our theory to synthetically generated QCQP problems. Our primary goal is to validate our theoretical guarantees for ||-`MRM` working in parameter-free fashion "out-of-the-box" and highlight a surprising disconnect where performance outscales our theory's predictions. Our implementation is not state-of-the-art, and so we restrict our attention to understanding ||-`MRM` rather than comparisons with other methods. We consider QCQPs of the form

$$p^* = \begin{cases} \max & f_0(x) := r_0 - q_0^T x - \frac{1}{2} x^T P_0 x \\ \text{s.t.} & f_j(x) := r_j - q_j^T x - \frac{1}{2} x^T P_j x \geq 0 \quad \forall j = 1, \ldots, m\ . \end{cases} \tag{5.1}$$

where the matrices P_j are symmetric and positive definite.

All our synthetic problems are constructed as follows. The matrices P_j take the form $P_j = G_j^T G_j + \lambda I$ for all $j = 0, 1, \ldots, m$, where $\lambda = 0.01$, $I \in \mathbb{R}^{n \times n}$ is the identity matrix, and each entry of $G_j \in \mathbb{R}^{n \times n}$ is sampled independently from the standard normal distribution. Each q_j is drawn independently from the normal distribution with mean 0 and covariance $\sigma_j I$. To avoid the trivial case where the solution is interior to the constraints, we take $\sigma_0 = 10$ and $\sigma_j = 1$ for $j \geq 1$. Finally, to guarantee a Slater point exists, we ensure $f_j(0) > 0$ by selecting r_j independently and uniformly from $[0.1, 1.1]$. In all cases, $\mathcal{E} = \mathbb{R}^{200}$ with the standard Euclidean norm. Code implementing these experiments can be found at `https://github.com/samaktbo/Parallel-MultiRadial-Method`.

5.1 Performance of MRM with Varied Subproblem Solvers

First, we investigate how the ||-`MRM` method performs under different first-order solvers. Specifically, for each `fom` $\in \{$`subgrad`, `smooth`, `genGrad`$\}$, and $m \in \{10, 100, 1000\}$, Figure 1 shows the relative optimality gap $\frac{p^* - f_0(y_i^{best})}{p^* - f_0(x_0)}$ varies as a function of real-time and the number of iterations. We initialize each method with $x_0 = 0$ and Algorithm 2 with $b = 4.0$ and $N = 16$ parallel instances. We select the centers in an ideal fashion, i.e., we set $e_j = -P_j^{-1} q_j$, for each $j = 0, 1, \ldots, m$. We see that for relatively small number of constraints ($m \leq 100$), the generalized gradient method far outperforms the theoretical per-iteration convergence rate of $O(1/\sqrt{\varepsilon})$. However, this method scales poorly since each iteration requires a QP solve from `Mosek`, completing about 30 iterations in 3000 seconds for $m = 1000$. On the other hand, the smoothing method and the subgradient method scale reasonably

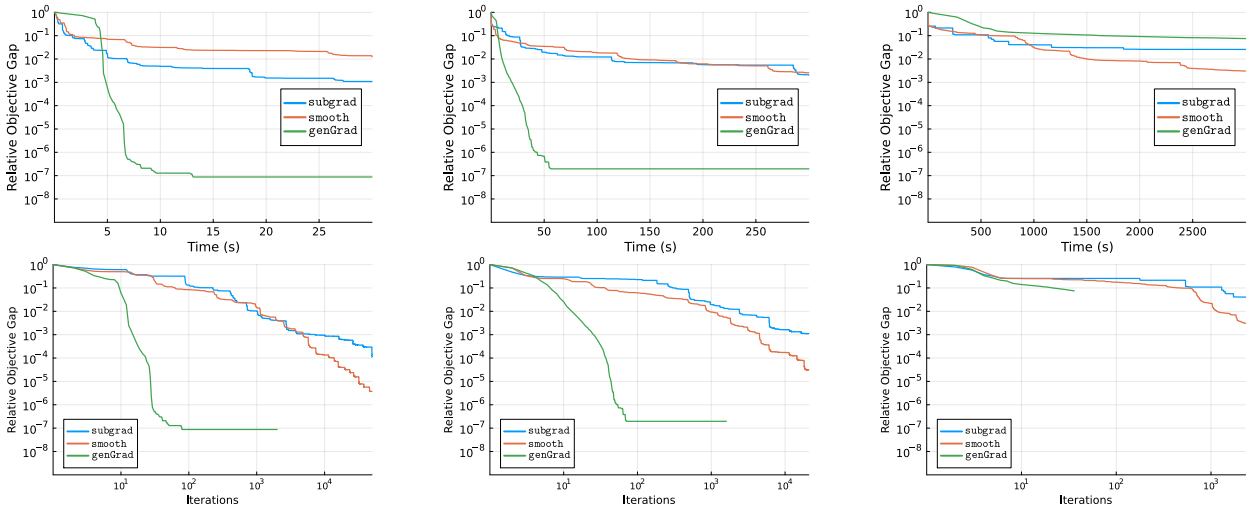

Figure 1: Performance of ||-MRM utilizing each of {subgrad, smooth, genGrad} in relative optimality gap $\frac{p^* - f_0(y_T^{best})}{p^* - f_0(x_0)}$, plotted against real-time and number of iterations. The number of constraints is $m = 10$, $m = 100$, and $m = 1000$, from left to right, respectively.

with m, even though their rate of convergence is slower. These method's convergence matches their theoretically predicted rates of $O(1/\varepsilon)$ and $O(1/\varepsilon^2)$, respectively, after slower convergence in the first hundred or so iterations, potentially corresponding to the $K_{\texttt{fom}}^{(l_0)}$ term in Theorem 4.1.

5.2 Effects of Multiradial Centers On Convergence

Next, we examine how the performance of Algorithm 2 is affected by the choice of centers $e_0, e_1, \ldots, e_m$ for problems of the form (5.1). We utilize the same set of underlying first-order methods and sample three QCQPs for the same selections of m as before. Then, for $K = 300$ target magnitudes of R ranging from 10^{-6} to 10^{-1}, we randomly sample centers e_j with controlled $R_{e_j}(S_j)$ (see full construction below). Surprisingly, Figure 2 shows the (relative) optimality gap of the iterates of Algorithm 2 reached is essentially independent of the choice of centers e_j and the related constant R. Practically, this indicates one need not spend much computational effort to find "good" centers to use. Conceptually, this indicates a gap between our theoretical bounds and actual performance. This is true for all three solvers and across problem sizes.

Note for a given e_j, computing $R_{e_j}(S_j)$ is a nontrivial nonconvex optimization problem. To avoid this difficulty, rather than randomly sampling e_j and computing the resulting R, we generate the e_j in such a way that $R_{e_j}(S_j)$ has a closed form. Namely for any x_j with $f_j(x_j) = 0$ and $0 < \alpha \leq 1$, $e_j = x_j + \frac{\alpha}{\|P_j\|}\nabla f_j(x_j)$ has $f_j(e_j) > 0$ and $R_{e_j}(S_j) = \alpha\frac{\|\nabla f_j(x_j)\|}{\|P_j\|}$ since f_j is $\|P_j\|$-smooth. For each of our $K = 300$ trials, we use this construction for e_j, setting α uniformly between 0.01 and 1 (in log-scale) and $x_j = \bar{e}_j + \sqrt{2f(\bar{e}_j)}P_j^{-\frac{1}{2}}u_j$ where $\bar{e}_j = -P_j^{-1}q_j$ and u_j is sampled uniformly from the unit sphere. The scaling $\sqrt{2f(\bar{e}_j)}$ ensures that $f(x_j) = 0$.

6 Deferred Analysis of Parallel MultiRadial Method

We begin by bounding the rate at which the dual gaps $1 - d(\tau_i^{best})$ decrease. Recall that $\texttt{fom}^{(l)}$ of Algorithm 2 restarts at $i \geq 0$ if $\tau_{i+1}^{best} \leq \frac{1}{1+\delta^{(l)}}\tau_i^{(l)}$ and $i_0^{(l)} < i_1^{(l)} < i_2^{(l)} \ldots$ denotes the sequence of

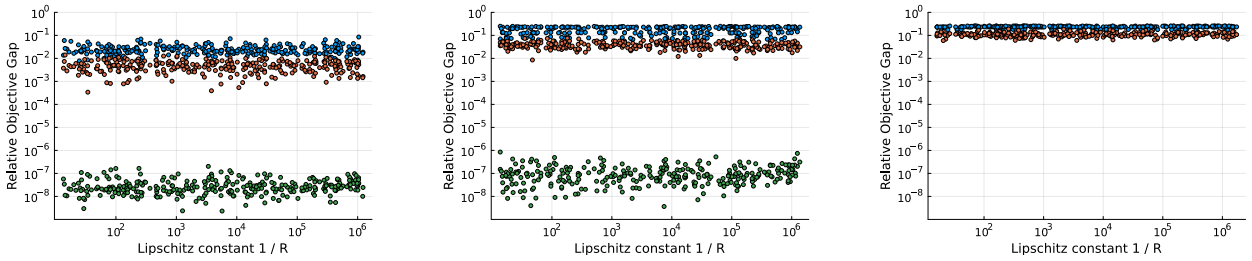

Figure 2: Final relative optimality gap $\frac{p\ -f_0(y_T^{best})}{p\ -f_0(x_0)}$ vs Lipschitz constant for three solvers: `subgrad` (blue), `smooth` (orange), and `genGrad` (green). The number of constraints is $m = 10$, $m = 100$, and $m = 1000$, from left to right. For $m = 1000$, `genGrad` was prohibitively costly to run.

iterations where the lth first-order method instance restarted.

Lemma 6.1. *Under Assumptions A - D, if* $\frac{b\delta^{(l)}}{\rho} \leq 1 - d(\tau_i^{best})$ *at iteration* $i \geq 0$ *of Algorithm 2 with* $b \geq 2$*, then*

$$1 - d(\tau_i^{best}) < \frac{b\delta^{(l)}}{\rho} \quad \textit{for all} \quad i' > i + \frac{1}{\log(1+\delta^{(l)})} \log\left(\frac{p(\tau_i^{(l)})}{1+b\delta^{(l)}}\right) K_{\texttt{fom}}^{(l)}.$$

Proof. It suffices to bound the number of iterations until $p(\tau_i^{best}) < 1 + b\delta^{(l)}$ holds since by Theorem 3.4, this implies $1 - d(\tau_i^{best}) < \frac{b\delta^{(l)}}{\rho}$. Let $i_k^{(l)}$ denote the first iteration after i where $\texttt{fom}^{(l)}$ restarts and $i_{\hat{k}}^{(l)}$ denote the first iteration with $p(\tau_{i_{\hat{k}}^{(l)}}^{(l)}) < 1 + b\delta^{(l)}$. The restarting condition of $\texttt{fom}^{(l)}$ ensures $p(\tau_{i_{j+1}^{(l)}}^{(l)}) \leq (1+\delta^{(l)})^{-1} p(\tau_{i_j^{(l)}}^{(l)})$. Therefore

$$p(\tau_{i_{\hat{k}}^{(l)}}^{best}) \leq p(\tau_{i_{\hat{k}}^{(l)}}^{(l)}) \leq (1+\delta^{(l)})^{-(\hat{k}-k)} p(\tau_i^{(l)}) \ .$$

Hence after $\frac{1}{\log(1+\delta^{(l)})} \log\left(\frac{p(\tau_i^{(l)})}{1+b\delta^{(l)}}\right)$ restarts of $\texttt{fom}^{(l)}$, every iteration i' must have $p(\tau_i^{best}) < 1 + b\delta^{(l)}$ and hence $1 - d(\tau_i^{best}) < \frac{b\delta^{(l)}}{\rho}$.

All that remains to bound the number of iterations between consecutive restarts of $\texttt{fom}^{(l)}$ by $K_{\texttt{fom}}^{(l)}$. Consider some pair of restart times $i_k^{(l)} < i_{k+1}^{(l)}$ with $i_k^{(l)} \geq i$. If some first-order method instance $l' \neq l$ at an iteration $i' \leq i_k^{(l)} + K_{\texttt{fom}}^{(l)}$, finds an iterate improving τ^{best} to be less than $\tau_{i_k^{(l)}}^{(l)}/(1+\delta^{(l)})$, then $\texttt{fom}^{(l)}$ will restart with $i_{k+1}^{(l)} \leq i_k^{(l)} + K_{\texttt{fom}}^{(l)}$. Otherwise, $\texttt{fom}^{(l)}$ proceeds without interruption from other processes for at least $K_{\texttt{fom}}^{(l)}$ iterations. Then, by definition, some $i' \leq i_k^{(l)} + K_{\texttt{fom}}^{(l)}$ has $y_i^{(l)}$ be a $\delta^{(l)}/\rho$-minimizer of $\Phi_{\tau_i^{(l)}}$. Since $\frac{\delta^{(l)}}{\rho} \leq (1 - d(\tau_i^{best}))/b \leq (1 - d(\tau_{i_k^{(l)}}^{(l)}))/b$, Corollary 3.1 implies $y_i^{(l)}$ is feasible and $\tau_{i_k^{(l)}}^{(l)} f(y_i^{(l)}) - 1 \geq (1 - 1/b)\rho(1 - d(\tau_{i_k^{(l)}}^{(l)})) \geq \delta^{(l)}$. Hence $1/f(y_i^{(l)}) \leq \frac{1}{1+\delta^{(l)}} \tau_{i_k^{(l)}}^{(l)}$ and so $i_{k+1}^{(l)} \leq i_k^{(l)} + K_{\texttt{fom}}^{(l)}$.

□

6.1 Proof of Theorem 4.1

From Lemma 6.1, we arrive at the following.

Theorem 6.1. *Suppose Assumptions A - D hold and let* fom *be given. Then, for all* $\bar{\varepsilon} > 0$, *Algorithm 2 with* fom, $b \geq 2$, *and* $N = \lceil \log_b(1/\bar{\varepsilon}) \rceil$ *has*

$$1 - d(\tau_i^{best}) \leq \frac{b\bar{\varepsilon}}{\rho} \quad \textit{for all} \quad i \geq \frac{\log(\tau_0 p^*)}{\log(1+\frac{\rho}{b^2})} K_{\texttt{fom}}^{(l_0)} + \frac{5b^2(1+\frac{\rho}{b^2})}{4\rho c_{\tau_0}} \sum_{l=l_0+1}^{N} K_{\texttt{fom}}^{(l)}$$

In fact, any such i *has* $1 - d(\tau_i^{best}) \leq \frac{b\delta^{(N)}}{\rho}$.

Proof. Note that $\frac{b\delta^{(N)}}{\rho} = \frac{b}{b^N\rho} \leq \frac{b\bar{\varepsilon}}{\rho}$ so that the second statement of the theorem implies the first. Now, if $1 - d(\tau_0) \leq \frac{b\delta^{(N)}}{\rho}$ then there is nothing to prove. We therefore assume for the rest of the proof that $\frac{b}{b^N\rho} < 1 - d(\tau_0) \leq 1$.

Considering the partition of $[0,1]$ given by $\left\{0, \frac{b}{b^N\rho}, \frac{b}{b^{N-1}\rho}, \ldots, \frac{b}{b^{l_0}\rho}, 1\right\}$, Lemma 6.1 gives a bound on how long $1 - d(\tau_i^{best})$ can remain in each sub-interval of $(\frac{b}{b^N\rho}, 1]$. We get the number of iterations needed to have $1 - d(\tau_i^{best}) \leq \frac{b}{b^N\rho}$ by summing the total number of iterations needed to move $1 - d(\tau_i^{best})$ out of each sub-interval of $(\frac{b}{b^N\rho}, 1]$.

Note that $\frac{\log(p(\tau_0))}{\log(1+\frac{\rho}{b^2})} K_{\texttt{fom}}^{(l_0)} \geq \frac{1}{\log(1+\delta^{(l_0)})} \log\left(\frac{p(\tau_0)}{1+b\delta^{(l_0)}}\right) K_{\texttt{fom}}^{(l_0)}$ since $\rho \leq b/b^{l_0-1}$ by the definition of l_0. Therefore, if $\frac{b\delta^{(l_0)}}{\rho} < 1 - d(\tau_0)$, then $1 - d(\tau_i^{best}) < \frac{b\delta^{(l_0)}}{\rho}$ for all $i > \frac{\log(p(\tau_0))}{\log(1+\frac{\rho}{b^2})} K_{\texttt{fom}}^{(l_0)}$, by Lemma 6.1.

Note that the restarting condition implies $\tau_i^{(l)} \leq (1+\delta^{(l)})\tau_i^{best}$. Thus, any i with $1 - d(\tau_i^{best}) \leq \frac{b^2\delta^{(l)}}{\rho}$ has

$$\begin{aligned} p(\tau_i^{(l)}) \leq (1+\delta_i^{(l)}) p(\tau_i^{best}) \leq 1 + \frac{1 - d(\tau_i^{best})}{c_{\tau_i^{best}}} + p(\tau_i^{best})\delta^{(l)} &\leq 1 + \left[\frac{b^2}{\rho c_{\tau_i^{best}}} + p(\tau_i^{best})\right]\delta^{(l)} \\ &\leq 1 + \left[\frac{b^2}{\rho c_{\tau_0}} + p(\tau_0)\right]\delta^{(l)}, \end{aligned}$$

where the second inequality is by Theorem 3.3, the third is by assumption, and the fourth holds because $\tau_i^{best} \leq \tau_0$. Therefore,

$$\begin{aligned} \frac{1}{\log(1+\delta^{(l)})} \log\left(\frac{p(\tau_i^{(l)})}{1+b\delta^{(l)}}\right) &= \frac{1}{\ln(1+\delta^{(l)})} \ln\left(1 + \frac{p(\tau_i^{(l)}) - 1 - b\delta^{(l)}}{1+b\delta^{(l)}}\right) \\ &\leq \frac{1}{\ln(1+\delta^{(l)})} \left(\frac{p(\tau_i^{(l)}) - 1 - b\delta^{(l)}}{1+b\delta^{(l)}}\right) \\ &\leq \frac{1}{\ln(1+\delta^{(l)})} \left[\frac{b^2}{\rho c_{\tau_0}} + p(\tau_0) - b\right] \frac{\delta^{(l)}}{1+b\delta^{(l)}} \\ &\leq \frac{\delta^{(l)}}{\ln(1+\delta^{(l)})} \frac{b^2}{\rho c_{\tau_0}} \left[1 + \rho \frac{c_{\tau_0}[p(\tau_0) - 1]}{b^2}\right] \\ &\leq \frac{5b^2}{4\rho c_{\tau_0}} \left(1 + \frac{\rho}{b^2}\right) \end{aligned}$$

where the first inequality follows from $\ln(1+x) \leq x$ and the last follows because $\frac{\delta}{\ln(1+\delta)} \leq \frac{5}{4}$ for any $\delta \in (0, 1/2]$ and $c_{\tau_0}[p(\tau_0) - 1] \leq 1$ by Theorem 3.3. As such, if $\frac{b\delta^{(l)}}{\rho} < 1 - d(\tau_i^{best}) \leq \frac{b^2\delta^{(l)}}{\rho}$ for $l > l_0$, then $1 - d(\tau_i^{best}) \leq \frac{b\delta^{(l)}}{\rho}$ for all $i' > i + \frac{5b^2(1+\frac{\rho}{b^2})}{4\rho c_{\tau_0}} K_{\texttt{fom}}^{(l)}$ by Lemma 6.1. Summing everything completes the proof. □

From this theorem, our originally claimed Theorem 4.1 follows directly: Given ε and i as in Theorem 4.1, consider $\bar{\varepsilon} = \frac{c_{\tau_0}\rho\varepsilon}{bp}$. Then the above result ensures

$$\begin{aligned}
\frac{b\bar{\varepsilon}}{\rho} &\geq 1 - d(\tau_i^{best}) && \text{[Theorem 6.1]} \\
&\geq c_{\tau_i^{best}}(\tau_i^{best}p^* - 1) && \text{[Theorem 3.3]} \\
&\geq c_{\tau_0}\frac{p^* - f(y_i^{best})}{f(y_i^{best})} && [c_{\tau_0} \leq c_{\tau_i^{best}}] \\
&\geq c_{\tau_0}\frac{p^* - f(y_i^{best})}{p^*}.
\end{aligned}$$

Since y_i^{best} is always feasible, the proof is complete.

Acknowledgements. This work was supported in part by the Air Force Office of Scientific Research under award number FA9550-23-1-0531.

References

[1] Michael P. Friedlander, Ives Macêdo, and Ting Kei Pong. Gauge optimization and duality. *SIAM Journal on Optimization*, 24(4):1999–2022, 2014.

[2] A. Y. Aravkin, J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, and K. J. MacPhee. Foundations of gauge and perspective duality. *SIAM Journal on Optimization*, 28(3):2406–2434, 2018.

[3] James Renegar. "Efficient" Subgradient Methods for General Convex Optimization. *SIAM Journal on Optimization*, 26(4):2649–2676, 2016.

[4] James Renegar. Accelerated first-order methods for hyperbolic programming. *Mathematical Programming*, 173(1-2):1–35, 2019.

[5] Benjamin Grimmer. Radial subgradient method. *SIAM Journal on Optimization*, 28(1):459–469, 2018.

[6] Benjamin Grimmer. Radial duality part ii: applications and algorithms. *Mathematical Programming*, 2023.

[7] Zakaria Mhammedi. Efficient projection-free online convex optimization with membership oracle. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5314–5390. PMLR, 02–05 Jul 2022.

[8] Zhou Lu, Nataly Brukhim, Paula Gradu, and Elad Hazan. Projection-free adaptive regret with membership oracles. In Shipra Agrawal and Francesco Orabona, editors, *International Conference on Algorithmic Learning Theory, February 20-23, 2023, Singapore*, volume 201 of *Proceedings of Machine Learning Research*, pages 1055–1073. PMLR, 2023.

[9] Ning Liu and Benjamin Grimmer. Gauges and accelerated optimization over smooth and/or strongly convex sets, 2023.

[10] Benjamin Grimmer. Radial duality part i: foundations. *Mathematical Programming*, 2023.

[11] B. T. Polyak. A general method of solving extremum problems. *Sov. Math., Dokl.*, 8:593–597, 1967.

[12] M.R. Metel and A. Takeda. Primal-dual subgradient method for constrained convex optimization problems. *Optimization Letters*, 15:1491–1504, 2021.

[13] Zhe Zhang and Guanghui Lan. Solving convex smooth function constrained optimization is almost as easy as unconstrained optimization, 2022.

[14] Robert M. Freund. Dual gauge programs, with applications to quadratic programming and the minimum-norm problem. *Mathematical Programming*, 38(1):47–67, 1987.

[15] David L. Applegate, Mateo D'iaz, Oliver Hinder, Haihao Lu, Miles Lubin, Brendan O'Donoghue, and Warren Schudy. Practical large-scale linear programming using primal-dual hybrid gradient. In *Neural Information Processing Systems*, 2021.

[16] Kinjal Basu, Amol Ghoting, Rahul Mazumder, and Yao Pan. ECLIPSE: An extreme-scale linear program solver for web-applications. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 704–714. PMLR, 13–18 Jul 2020.

[17] Qi Deng, Qing Feng, Wenzhi Gao, Dongdong Ge, Bo Jiang, Yuntian Jiang, Jingsong Liu, Tianhao Liu, Chenyu Xue, Yinyu Ye, and Chuwen Zhang. New developments of admm-based interior point methods for linear programming and conic programming, 2023.

[18] Yinyu Ye Tianyi Lin, Shiqian Ma and Shuzhong Zhang. An admm-based interior-point method for large-scale linear programming. *Optimization Methods and Software*, 36(2-3):389–424, 2021.

[19] Haihao Lu and Jinwen Yang. A practical and optimal first-order method for large-scale convex quadratic programming, 2023.

[20] Yurii Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[21] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM Journal on Optimization*, 22:557–580, 2012.

[22] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014.

[23] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152:381–404, 2015.

[24] James Renegar and Benjamin Grimmer. A simple nearly optimal restart scheme for speeding up first-order methods. *Foundations of Computational Mathematics*, 22:211–256, 2022.

A Proofs of Corollaries

Throughout this appendix, we let

$$D := \max\{D(S_j) \mid j = 0, 1, \ldots, m\} \quad \text{and} \quad R := \min\{R_{e_j}(S_j) \mid j = 0, 1, \ldots, m\},$$

where $e_1, \ldots, e_m$ are the reference points defining $\gamma_{S_1,e_1}, \ldots, \gamma_{S_m,e_m}$.

A.1 Proof of Corollary 4.1

Let $\bar{\varepsilon} = \frac{c_{\tau_0}\rho\varepsilon}{bp}$ and $\tilde{N} = \lceil \log_b(\frac{bp}{c_{\tau_0}\rho\varepsilon}) \rceil$. By Theorem 4.1, it suffices to show that $K^{(l_0)}_{\texttt{subgrad}}$ and $\sum_{l=l_0+1}^{\tilde{N}} K^{(l)}_{\texttt{subgrad}}$ are bounded by $O(1/\varepsilon^2)$. Since Φ_τ is $1/R$-Lipschitz for all $\tau > 0$, it follows that $K^{(l)}_{\texttt{subgrad}} \le \frac{(D_0/R)^2}{(\delta^{(l)}/\rho)^2} = (\rho b^l D_0/R)^2$. Now, we have $b^{l_0} \le b^2/\rho$ by definition, hence $K^{(l_0)}_{\texttt{subgrad}} \le b^4 \frac{D_0^2}{R^2}$ is constant with respect to ε. For $l > l_0$, we have $b^l = \frac{b \cdot b^{\tilde{N}-1}}{b^{\tilde{N}-l}} \le \frac{b}{\bar{\varepsilon} b^{\tilde{N}-l}} = \frac{b}{\rho}\left(\frac{bp}{c_{\tau_0}}\right)\frac{1}{b^{\tilde{N}-l}}\frac{1}{\varepsilon}$. Therefore

$$\sum_{l=l_0+1}^{\tilde{N}} K^{(l)}_{\texttt{subgrad}} \le \left[b^2 \left(\frac{bp^*}{c_{\tau_0}}\right)^2 \sum_{l=l_0+1}^{\tilde{N}} \frac{1}{b^{2(\tilde{N}-l)}} \right] \frac{D_0^2}{R^2}\frac{1}{\varepsilon^2} = O\left(\frac{1}{\varepsilon^2}\right).$$

A.2 Proof of Corollary 4.2

Let $\bar{\varepsilon} = \frac{c_{\tau_0}\rho\varepsilon}{bp}$, $\tilde{N} = \lceil \log_b(\frac{bp}{c_{\tau_0}\rho\varepsilon})\rceil$, and $\theta^{(l)} = \frac{\delta^{(l)}}{2\log(m+1)}$. By Theorem 4.1, it suffices to show that $K^{(l_0)}_{\texttt{smooth}}$ and $\sum_{l=l_0+1}^{\tilde{N}} K^{(l)}_{\texttt{smooth}}$ are bounded by $O(1/\varepsilon)$. Recall that $\texttt{smooth}^{(l)}$ corresponds to Nesterov's accelerated method applied to the smoothed objective

$$\Phi_{\tau_i^{(l)},\theta^{(l)}}(y) := \theta^{(l)} \log\left(\exp\left(\frac{(\tau_i^{(l)} f)^{\Gamma,e_0}}{\theta^{(l)}}\right) + \sum_{j=1}^{m} \exp\left(\frac{\varphi_{S_j}(y)}{\theta^{(l)}}\right)\right).$$

First, we observe that all of the components $(\tau_i^{(l)} f)^{\Gamma,e_0}, \varphi_{S_1}, \ldots, \varphi_{S_m}$ are all L-smooth and M-Lipschitz where $M = 1/R$ and $L = \max\{(1+\frac{D_0}{R_0})^3\tau_0\beta, \frac{R+\beta D^2}{R^3}\}$. The smoothness and Lipschitz continuity of each identifier is verified below in Lemma B.1. The $(1+\frac{D_0}{R_0})^3\tau_0\beta$-smoothness of $(\tau_i^{(l)} f)^{\Gamma,e_0}$ follows from [6, Corollary 1] and noting $\tau_i^{(l)} \le \tau_0$. The $1/R$-Lipschitz continuity of $(\tau_i^{(l)} f)^{\Gamma,e_0}$ follows from (2.10). From these bounds, it follows that $\Phi_{\tau_i^{(l)},\theta^{(l)}}$ is $\max\{(1+\frac{R_0}{D_0})^3\tau_0\beta, \frac{R+\beta D^2}{R^3}\} + \frac{M^2}{\theta^{(l)}}$-smooth (see Appendix B of [21]). Hence

$$\begin{aligned} K_{\texttt{smooth}} &\le 2\sqrt{\frac{2LD^2}{\delta^{(l)}/\rho} + \frac{4M^2D^2\log(m+1)}{(\delta^{(l)}/\rho)^2}} \\ &\le 2\sqrt{\frac{2LD^2}{b} + 4M^2\log(m+1)D^2}\ \frac{\rho}{\delta^{(l)}} \end{aligned}$$

where the second inequality uses that all $l \ge l_0$ have $\frac{\delta^{(l)}}{\rho} < 1/b$.

From this, it follows that $K^{(l_0)}_{\texttt{smooth}} \le 2\sqrt{\frac{2LD^2}{b} + 4M^2\log(m+1)D^2}\ b^2$ is constant with respect to ε as $1/\delta^{(l_0)} \le b^2/\rho$. For $l > l_0$, we have $1/\delta^{(l)} = b^l \le \frac{b}{\rho}\left(\frac{bp}{c_{\tau_0}}\right)\frac{1}{b^{\tilde{N}-l}}\frac{1}{\varepsilon}$. Therefore, the total iteration bound of Theorem 4.1 scales with ε as

$$\sum_{l=l_0+1}^{\tilde{N}} K^{(l)}_{\texttt{smooth}} \le 2\sqrt{\frac{2LD^2}{b} + 4M^2\log(m+1)D^2}\left[b\left(\frac{bp^*}{c_{\tau_0}}\right)\sum_{l=0}^{\infty} 1/b^l\right]\frac{1}{\varepsilon} = O(1/\varepsilon).$$

A.3 Proof of Corollary 4.3

Note that $\gamma^2_{S_1,e_1}, \ldots, \gamma^2_{S_m,e_m}$ are all $2\frac{R+\beta D^2}{R^3}$-smooth by [9, Corollary 3.2]. In addition, $(\tau_i^{(l)} f)^{\Gamma,e_0}$ is $(1+\frac{R_0}{D_0})^3\tau_i^{(l)}\beta$-smooth by [6, Corollary 1]. Noting $\tau_i^{(l)} \le \tau_0$, it follows that $(\tau_i^{(l)} f)^{\Gamma,e_0}, \gamma^2_{S_1,e_1}, \ldots, \gamma^2_{S_m,e_m}$ are all $L = \max\{(1+\frac{D_0}{R_0})^3\tau_0\beta, \frac{2(R+\beta D^2)}{R^3}\}$-smooth. Therefore, Theorem 2.3.5 of [22] ensures that $K^{(l)}_{\texttt{genGrad}} \le 2\sqrt{LD_0^2}\cdot\sqrt{\frac{\rho}{\delta^{(l)}}}$.

Since $1/\delta^{(l_0)} \le b^2/\rho$, it follows that $K^{(l_0)}_{\texttt{genGrad}} \le 2\sqrt{LD_0^2}\ b$ is constant with respect to ε. Now, let $\bar{\varepsilon} = \frac{c_{\tau_0}\rho\varepsilon}{bp}$ and $\tilde{N} = \lceil \log_b(1/\bar{\varepsilon})\rceil$. For $l > l_0$, we have $1/\delta^{(l)} = b^l \le \frac{b}{\rho}\left(\frac{bp}{c_{\tau_0}}\right)\frac{1}{b^{\tilde{N}-l}}\frac{1}{\varepsilon}$. Therefore, the total iteration bound of Theorem 4.1 scales with ε as

$$\sum_{l=l_0+1}^{\tilde{N}} K^{(l)}_{\texttt{genGrad}} \le 2\sqrt{LD_0^2}\left[\sqrt{b}\left(\frac{bp^*}{c_{\tau 0}}\right)^{\frac{1}{2}}\sum_{l=0}^{\infty}(\sqrt{b})^{-l}\right]\frac{1}{\sqrt{\varepsilon}} = O(1/\sqrt{\varepsilon}).$$

B Smoothness of the identifiers in Corollary 4.2

Lemma B.1. *Let S be convex and $e \in \operatorname{int} S$. Then $\varphi_S : \mathcal{E} \to \mathbb{R}$, defined by*

$$\varphi_S(x) := \begin{cases} \gamma_{S,e}(x) & \text{if } \gamma_{S,e}(x) > 1 \\ \frac{1}{2}\gamma_{S,e}^2(x) + \frac{1}{2} & \text{otherwise} \end{cases},$$

is convex and $1/R_e(S)$-Lipschitz. If $\frac{1}{2}\gamma_{S,e}^2$ is L-smooth, φ_S is L-smooth. In particular, if S is β-smooth, then φ_S is $\frac{R_e(S)+\beta D_e(S)^2}{R_e(S)^3}$-smooth for $D_e(S) := \sup\{\|x - e\| \mid x \in S\}$.

Proof. We will show that $\varphi_S(x) = \frac{1}{2} + \inf_{y\in\mathcal{E}} \gamma_{S,e}(y) + \frac{1}{2}\gamma_{S,e}^2(e+x-y)$. Note that this will immediately impy φ_S is convex and as smooth as $\frac{1}{2}\gamma_{S,e}^2$ because infimal convolutions preserve convexity and smoothness. Lipschitz continuity is also immediate as the subgradients of φ_S and $\gamma_{S,e}$ are bounded above by $1/R_e(S)$.

It remains to show $\varphi_S(x) = \frac{1}{2} + \inf_{y\in\mathcal{E}} \gamma_{S,e}(y) + \frac{1}{2}\gamma_{S,e}^2(e + x - y)$. We prove that for any $\mu > 0$,

$$\inf_{y\in\mathcal{E}} \gamma_{S,e}(y) + \frac{1}{2\mu}\gamma_{S,e}^2(e + x - y) = \begin{cases} \gamma_{S,e}(x) - \frac{\mu}{2} & \text{if } \gamma_{S,e}(x) > \mu \\ \frac{1}{2\mu}\gamma_{S,e}^2(x) & \text{otherwise.} \end{cases} \tag{B.1}$$

By the sum rule and chain rule, which apply as $\gamma_{S,e}$ and $\gamma_{S,e}^2$ are convex and real-valued, the necessary and sufficient condition to attain the infimum above is

$$0 \in \partial\gamma_{S,e}(y) - \frac{1}{\mu}\gamma_{S,e}(e + x - y)\partial\gamma_{S,e}(e + x - y).$$

The result in (B.1) holds because $y = e$ or $y = e + \left[1 - \frac{\mu}{\gamma_{S,e}(x)}\right](x - e)$ satisfies the sufficient condition accordingly as $\gamma_{S,e}(x) \le \mu$ or $\gamma_{S,e}(x) > \mu$. Indeed, since $0 \in \partial\gamma_{S,e}(e)$ and $\partial\gamma_{S,e}(x) \subset \partial\gamma_{S,e}(e)$ for any x, it follows from convexity of the subdifferential that $\frac{1}{\mu}\gamma_{S,e}(x)\partial\gamma_{S,e}(x) \subset \partial\gamma_{S,e}(e)$ if $\gamma_{S,e}(x) \le \mu$. Therefore, $y = e$ attains the infimum if $\gamma_{S,e}(x) \le \mu$.

Now suppose $\gamma_{S,e}(x) > \mu$ and let $y = e + \left[1 - \frac{\mu}{\gamma_{S,e}(x)}\right](x - e)$. Recall that for any $\alpha > 0$, we have $\gamma_{S,e}(e+\alpha(x-e)) = \alpha\gamma_{S,e}(x)$ and $\partial\gamma_{S,e}(e+\alpha(x-e)) = \partial\gamma_{S,e}(x)$. Noting that $e+x-y = e+\frac{\mu}{\gamma_{S,e}(x)}(x-e)$, we conclude that

$$\gamma_{S,e}(e + x - y) = \frac{\mu}{\gamma_{S,e}(x)}\gamma_{S,e}(x) = \mu \quad \text{and} \quad \partial\gamma_{S,e}(y) = \partial\gamma_{S,e}(x) = \partial\gamma_{S,e}(e + x - y).$$

So, $\partial\gamma_{S,e}(y) = \frac{1}{\mu}\gamma_{S,e}(e + x - y)\partial\gamma_{S,e}(e + x - y)$, and y satisfies the sufficient condition.

Plugging in the suitable minimizer $y = e$ or $y = e + \left[1 - \frac{\mu}{\gamma_{S,e}(x)}\right](x - e)$ accordingly into $\gamma_{S,e}(y) + \frac{1}{2\mu}\gamma_{S,e}^2(e + x - y)$ gives (B.1). Finally, if S is β-smooth and $D_e(S) < \infty$, then $\frac{1}{2}\gamma_{S,e}^2$ is $\frac{R_e(S)+\beta D_e(S)^2}{R_e(S)^3}$-smooth by [9, Corollary 3.2]. Therefore, φ_S is smooth with the same smoothness constant. □