Radial Duality
Part II: Applications and Algorithms

Benjamin Grimmer*

Abstract

The first part of this work established the foundations of a radial duality between nonnegative optimization problems, inspired by the work of Renegar [38]. Here we utilize our radial duality theory to design and analyze projection-free optimization algorithms that operate by solving a radially dual problem. In particular, we consider radial subgradient, smoothing, and accelerated methods that are capable of solving a range of constrained convex and nonconvex optimization problems and that can scale-up more efficiently than their classic counterparts. These algorithms enjoy the same benefits as their predecessors, avoiding Lipschitz continuity assumptions and costly orthogonal projections, in our newfound, broader context. Our radial duality further allows us to understand the effects and benefits of smoothness and growth conditions on the radial dual and consequently on our radial algorithms.

1 Introduction

The first part of this work [17] established a theory of radial duality relating nonnegative optimization problems through a projective transformation, extending the ideas of Renegar [38] from their origins in conic programming. We give a minimal overview here of our radial duality theory needed to begin algorithmically benefiting from it and then a fuller but terse summary in Section 2.3 of the core results necessary to derive our radial optimization guarantees.

For a finite dimensional Euclidean space $\mathcal{E}$, our three transformations of interest are the radial point transformation, radial set transformation, and upper radial function transformation, which are denoted by

$$\Gamma(x,u) = (x,1)/u,$$

$$\Gamma S = \{\Gamma(x,u) \mid (x,u) \in S\},$$

$$f^{\Gamma}(y) = \sup\{v > 0 \mid (y,v) \in \Gamma(\text{epi } f)\}$$

for any point $(x,u) \in \mathcal{E} \times \mathbb{R}_{++}$, set $S \subseteq \mathcal{E} \times \mathbb{R}_{++}$, and function $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, respectively. Here $\overline{\mathbb{R}}_{++}$ denotes the extended positive reals $\mathbb{R}_{++} \cup \{0, +\infty\}$. It is immediate that the point and set transformations are dual since

$$\Gamma\Gamma(x,u) = \Gamma\frac{(x,1)}{u} = \frac{(x/u,1)}{1/u} = (x,u).$$

*School of Operations Research and Information Engineering, Cornell University, Ithaca, NY 14850, USA; people.orie.cornell.edu/bdg79/

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1650441. This work was partially done while the author was visiting the Simons Institute for the Theory of Computing. It was partially supported by the DIMACS/Simons Collaboration on Bridging Continuous and Discrete Optimization through NSF grant #CCF-1740425.

Central to establishing our theory of radial duality is the characterization of exactly when this duality carries over to the function transformation. We say a function f is *upper radial* if the perspective function $f^p(y,v) = v \cdot f(y/v)$ is upper semicontinuous and nondecreasing in $v \in \mathbb{R}_{++}$. Moreover, it is *strictly upper radial* if it is strictly increasing in v whenever $f^p(y,v) \in \mathbb{R}_{++}$. The cornerstone theorem of our radial duality [17, Theorem 3.2] is that

$$f = f^{\Gamma\Gamma} \iff f \text{ is upper radial.} \tag{1}$$

The duality of the radial function transformation provides a duality between optimization problems. For any strictly upper radial function $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, consider the primal problem

$$p^* = \max_{x\in\mathcal{E}} f(x). \tag{2}$$

Then the radially dual problem is given by

$$d^* = \min_{y\in\mathcal{E}} f^\Gamma(y) \tag{3}$$

and has $(\operatorname{argmax} f) \times \{p^*\} = \Gamma\big((\operatorname{argmin} f^\Gamma) \times \{d^*\}\big)$. Thus maximizing f is equivalent to minimizing f^Γ and solutions can be converted between these problems by applying the radial point transformation Γ or its inverse (which is also Γ by duality).

Importantly, the two nonnegative optimization problems (2) and (3) can exhibit very different structural properties. For example, consider maximizing $f(x) = \sqrt{1 - \|x\|^2_{2+}}$ which takes value zero outside the unit ball and has arbitrarily large gradients and Hessians as x approaches the boundary of this ball. Its radial dual $f^\Gamma(y) = \sqrt{1 + \|y\|_2^2}$ has full domain with gradients and Hessians bounded in norm by one everywhere. Thus our radial duality theory poses an opportunity to extend the reach of many standard optimization algorithms reliant on such structure. The previous works of Renegar [38] and Grimmer [16] analyzing subgradient methods and Renegar [39] employing accelerated smoothing techniques on a radial reformulation of the objective critically rely on the reformulation being uniformly Lipschitz continuous, which always occurs in the special cases of the radial dual that they consider.

Our Contributions. This work leverages our radial duality theory to present and analyze projection-free radial optimization algorithms in this newfound, wider context than previous works were able to. Finding that a mild condition ensures the radial dual is uniformly Lipschitz continuous, we analyze a radial subgradient method for a broad range of non-Lipschitz primal problems with or without concavity. Observing that constraints radially transform into related gauges, we propose a radial smoothing method that takes advantage of this structure for concave maximization. Further, we find that our radial transformation extends smoothness on a level set of the primal to hold globally in the radial dual, which prompts our analysis of a radial accelerated method. Of greater importance than these particular algorithms, this work aims to demonstrate the breadth of applications and algorithms that can be approached using our radial duality theory.

Outline. We begin with a motivating example of the computational benefits and scalability that follow from designing algorithms based on the radial dual (3) in Section 2. Then Section 3 formally establishes algorithmically useful properties of our radial dual, namely Lipschitz continuity, smoothness, and growth conditions. Finally, Section 4 addresses the convergence of our radial algorithms for concave maximization and Section 5 addresses applications and guarantees in nonconcave maximization.

2 A Motivating Setting of Polyhedral Constraints

We begin by motivating the algorithmic usefulness of our radial duality by considering optimization with polyhedral constraints. Consider any maximization problem with upper semicontinuous objective $f\colon \mathbb{R}^n \to \mathbb{R}\cup\{-\infty\}$ and m inequality constraints $a_i^T x \leq b_i$ given by

$$\begin{cases} \max_x & f(x) \\ \text{s.t.} & Ax \leq b. \end{cases} \tag{4}$$

We assume this problem is feasible. Then without loss of generality, we have $0 \in \operatorname{int}(\{x \mid Ax \leq b\} \cap \{x \mid f(x) > 0\})$. This can be achieved by computing any point x_0 in the relative interior of $\{x \mid Ax \leq b\} \cap \{x \mid f(x) \in \mathbb{R}\}$ and then (i) translating the problem to place x_0 at the origin, (ii) adding a constant to the objective to ensure $f(0) > 0$, and (iii) if needed, re-parameterizing the problem[1] to only consider the smallest subspace containing $\{x \mid Ax \leq b\}\cap\{x \mid f(x) > 0\}$. Note that doing this translation suffices to guarantee that any concave f will have $f_+(x) := \max\{f(x), 0\}$ be strictly upper radial by [17, Proposition 3.8]. We will only make the weaker assumption here that f_+ is strictly upper radial rather than the narrower case of it being concave. Then this problem can be reformulated as the following nonnegative optimization problem of our primal form (2)

$$\begin{cases} \max_x & f_+(x) \\ \text{s.t.} & Ax \leq b \end{cases} = \max_x \min_i \Big\{ f_+(x),\ \iota_{a_i^T x \leq b_i}(x) \Big\}$$

where $\iota_{a_i^T x\leq b_i}(x) = \begin{cases} +\infty & \text{if } a_i^T x \leq b_i \\ 0 & \text{if } a_i^T x > b_i \end{cases}$ is an indicator function for each inequality constraint. Note that each $\iota_{a_i^T x \leq b_i}$ is strictly upper radial since 0 is strictly feasible and so applying [17, Proposition 3.9] ensures the primal objective $\min_i \Big\{ f_+(x),\ \iota_{a_i^T x\leq b_i}(x)\Big\}$ is strictly upper radial. Then we can compute the radially dual optimization problem (3) using [17, Proposition 3.10] as

$$\min_y \max_i \big\{ f_+^\Gamma(y),\ a_i^T y / b_i \big\} \tag{5}$$

since the radial transformation of each indicator function is linear

$$\begin{aligned} \iota^\Gamma_{a_i^T x \leq b_i}(y) &= \sup\Big\{ v > 0 \mid v \cdot \iota_{a_i^T x\leq b_i}(y/v) \leq 1 \Big\} \\ &= \sup\big\{ v > 0 \mid a_i^T (y/v) > b_i \big\} \\ &= (a_i^T y / b_i)_+. \end{aligned}$$

We drop the nonnegative thresholding on each $a_i^T y/b_i$ above since $f_+^\Gamma(y)$ is nonnegative.

Importantly, the dual formulation (5) is unconstrained, unlike the primal, since the primal inequality constraints have transformed into simple linear lower bounds on the radially dual objective. This dual further profits from the structure of its objective function as it is often globally Lipschitz continuous (a common property among radial duals that we will show in Proposition 3.1) and has the simple form of a finite maximum. This radially dual structure gives us an algorithmic angle of attack not available in the primal problem.

[1] Instead of using a re-parameterization, one can explicitly include equality constraints in our model. The details of this approach are given in Section 2.2.1, where we see that equality constraints are unaffected by the radial dual.

2.1 Quadratic Programming

To make these benefits concrete, consider solving a generic quadratic program of the following form

$$\begin{cases} \max_x & 1 - \frac{1}{2}x^TQx - c^Tx \\ \text{s.t.} & Ax \leq b \end{cases} \tag{6}$$

for some $Q \in \mathbb{R}^{n\times n}$, $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, and $b \in \mathbb{R}^m_{++}$. We reformulate this problem as the following nonnegative optimization problem of the form (2)

$$\begin{cases} \max_x & 1 - \frac{1}{2}x^TQx - c^Tx \\ \text{s.t.} & Ax \leq b \end{cases} = \max_x \min_i \left\{ (1 - \frac{1}{2}x^TQx - c^Tx)_+,\ \iota_{a_i^Tx\leq b_i}(x) \right\}.$$

Whenever this primal objective is strictly upper radial, the radial dual of our quadratic program is[2]

$$\min_y \max_i \left\{ \left(\frac{c^Ty + 1 + \sqrt{(c^Ty+1)^2 + 2y^TQy}}{2} \right)_+,\ a_i^Ty/b_i \right\} \tag{7}$$

where the first term in our maximum is set to zero if $(c^Ty+1)^2 + 2y^TQy < 0$ as can occur for nonconcave primal objectives. We find that our radial duality holds here whenever $\frac{1}{2}x^TQx > -1$ for all $Ax \leq b$. This captures two natural settings: (i) when the primal objective is concave (as Q is positive semidefinite) or (ii) when the primal objective is nonconcave but has a compact feasible region (since we can rescale the objective to be $1 - \lambda x^TQx/2 - \lambda c^Tx$ without changing the set of maximizers but ensuring $\frac{\lambda}{2}x^TQx > -1$ everywhere). Section 5.1 shows more generally that any differentiable objective with compact constraints can be rescaled to apply our radial duality theory.

We verify that our primal objective is strictly upper radial (and so our radial duality holds) for this upper semicontinuous objective by checking when $f^p(y,\cdot)$ is strictly increasing on its domain. The partial derivative with respect to v of the perspective function

$$v \cdot \min_i \left\{ (1 - \frac{1}{2}(y/v)^TQ(y/v) - c^T(y/v))_+,\ \iota_{a_i^Tx\leq b_i}(y/v) \right\} = \begin{cases} v\left(1 - \frac{y^TQy}{2v^2} - \frac{c^Ty}{v}\right) & \text{if } A(y/v) \leq b \\ 0 & \text{otherwise} \end{cases}$$

is $1 + \frac{y^TQy}{2v^2}$ at every feasible y/v. This is always positive (and hence the perspective function is increasing in v) exactly when every $x = y/v$ with $Ax \leq b$ has $\frac{1}{2}x^TQx > -1$.

2.1.1. Quadratic Programming Numerics As previously noted, the radially dual formulation (7) is unconstrained and Lipschitz continuous despite the primal possessing neither of these properties. This differs from the structure found from taking a Lagrange dual [11] or gauge dual [14]. As a result, our radial dual is well set up for the application of a subgradient method. We consider the following *radial subgradient method* with stepsizes $\alpha_k > 0$ defined by Algorithm 1.

[2] Our calculation of the radial dual of the quadratic objective $(1 - \frac{1}{2}x^TQx - c^Tx)_+$ follows by definition as

$$(1 - \frac{1}{2}x^TQx - c^Tx)_+^\Gamma(y) = \sup\left\{ v > 0 \mid v\left(1 - \frac{y^TQy}{2v^2} - \frac{c^Ty}{v}\right) \leq 1 \right\} = \sup\{v > 0 \mid v^2 - \frac{1}{2}y^TQy - (c^Ty+1)v \leq 0\}$$
$$= \left(\frac{c^Ty + 1 + \sqrt{(c^Ty+1)^2 + 2y^TQy}}{2} \right)_+.$$

Algorithm 1 The Radial Subgradient Method

Require: $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, $x_0 \in \text{dom } f$, $T \geq 0$

$(y_0, v_0) = \Gamma(x_0, f(x_0))$ *Transform into the radial dual*
for $k = 0 \ldots T-1$ **do**
  $y_{k+1} = y_k - \alpha_k \zeta'_k$, where $\zeta'_k \in \partial_P f^\Gamma(y_k)$ *Run the subgradient method*
end for
$(x_T, u_T) = \Gamma(y_T, f^\Gamma(y_T))$ *Transform back to the primal*
return x_T

Further noting that the radially dual problem is a finite maximum of simple smooth Lipschitz functions, we can apply the smoothing ideas of Nesterov [33]. Perhaps the most clear description of these techniques is given by Beck and Teboulle [3]. In particular, for any fixed $\eta > 0$, we consider the smooth function given by taking a "soft-max"

$$g_\eta(y) = \eta \log\left(\exp\left(\frac{c^Ty + 1 + \sqrt{(c^Ty+1)^2 + 2y^TQy}}{2\eta}\right) + \sum_{i=1}^{m} \exp\left(\frac{a_i^Ty}{b_i\eta}\right)\right) \tag{8}$$

which approaches our radially dual objective as $\eta \to 0$. Then we can minimize the radial dual up to accuracy $O(\eta)$ by minimizing this smoothed objective. Doing so with Nesterov's accelerated method gives the following *radial smoothing method* defined by Algorithm 2 (a similar radial algorithm was employed by Renegar [39] showing that the transformation of any hyperbolic programming problem also admits a smoothing that can be efficiently minimized).

Algorithm 2 The Radial Smoothing Method

Require: $f\colon \mathcal{E} \to \overline{\mathbb{R}}_{++}$, $x_0 \in \text{dom } f$, $\eta > 0$, $L_\eta > 0$, $T \geq 0$

$(y_0, v_0) = \Gamma(x_0, f(x_0))$ and $\tilde{y}_0 = y_0$ *Transform into the radial dual*
Let $g_\eta(y)$ denote an η-smoothing of $f^\Gamma(y)$
for $k = 0 \ldots T-1$ **do**
  $\tilde{y}_{k+1} = y_k - \nabla g_\eta(y_k)/L_\eta$ *Run the accelerated method*
  $y_{k+1} = \tilde{y}_{k+1} + \frac{k-1}{k+2}(\tilde{y}_{k+1} - \tilde{y}_k)$
end for
$(x_T, u_T) = \Gamma(y_T, f^\Gamma(y_T))$ *Transform back to the primal*
return x_T

The per iteration cost of these radial methods is controlled by the cost of evaluating one subgradient of the radially dual objective (7) or one gradient of our smoothing of the radially dual objective (8). Both of these can be done efficiently in closed form in terms of the two matrix-vector products Ay and Qy. Despite this low iteration cost, a feasible primal solution $(x_k, u_k) = \Gamma(y_k, f^\Gamma(y_k))$ is known at every iteration. Convergence guarantees for the radial subgradient and smoothing methods for concave maximization are given later in Sections 4.1 and 4.2.

Classic optimization algorithms that preserve feasibility at every iteration tend to have much higher iteration costs. Here we compare with three of the most standard first-order methods that enforce feasibility: projected gradient descent (or rather, projected gradient ascent)

$$x_{k+1} = \text{proj}_{\{x \mid Ax \leq b\}}(x + \nabla f(x)/L),$$

an accelerated projected gradient method

$$\begin{cases} \tilde{x}_{k+1} &= \text{proj}_{\{x \mid Ax \le b\}}(x_k + \nabla f(x_k)/L) \\ x_{k+1} &= \tilde{x}_{k+1} + \frac{k-1}{k+2}(\tilde{x}_{k+1} - \tilde{x}_k), \end{cases}$$

and the Frank-Wolfe method[3] with stepsize sequence $\beta_k > 0$

$$\begin{cases} \tilde{x}_{k+1} &\in \text{argmax}_x \left\{ \nabla f(x_k)^T x \mid Ax \le b \right\} \\ x_{k+1} &= x_k + \beta_k(\tilde{x}_{k+1} - x_k). \end{cases}$$

All three of these methods require solving a subproblem at each iteration. The projected gradient and accelerated gradient methods require repeated projection onto the polyhedron $\{x \mid Ax \le b\}$, which is itself an instance of (6) specialized to $Q = I$. The Frank-Wolfe method requires repeatedly solving a linear program over this polyhedron. Both of these operations are far more expensive than the matrix-vector products required by the radial subgradient and smoothing methods but may allow them to have a greater improvement in objective value per iteration[4].

To weigh this tradeoff, we consider running these five algorithms on synthetic quadratic programs given by drawing two matrices $A \in \mathbb{R}^{m \times n}$ and $P \in \mathbb{R}^{n \times 100}$ and a vector $c \in \mathbb{R}^n$ with i.i.d. Guassian entries and setting $Q = PP^T$ and all $b_i = 1$. Then we run each algorithm for 30 minutes on instances of size $(n, m) \in \{(100, 400), (400, 1600), (1600, 6400)\}$. Our numerical experiments are conducted on a four-core Intel i7-6700 CPU using Julia 1.4.1 and Gurobi 9.0.3 to solve any subproblems[5]. For each method, we set $x_0 = 0$ and use the following choice of stepsizes: the projected and accelerated gradient methods use $L = \lambda_{max}(Q)$, the Frank-Wolfe method uses an exact linesearch $\beta_k = \min\left(\frac{\nabla f(x_k)^T(\tilde{x}_{k+1} - x_k)}{\|P^T(\tilde{x}_{k+1} - x_k)\|^2}, 1\right)$, the radial subgradient method uses the Polyak stepsize $\alpha_k = \frac{f^\Gamma(y_k) - d^*}{\|\zeta'_k\|^2}$, and the radial smoothing method fixes $L_\eta = 0.1 \max\{\|a_i/b_i\|^2\}/\eta$ and $\eta \in \{10^{-8}, 10^{-8}, 10^{-6}\}$ for each of our three problem sizes.

The best primal objective value seen by each method is shown in realtime in Figure 1. First, we remark on the total number of iterations completed by each method in the allotted half hour, shown in the following table.

	$(n, m) = (100, 400)$	$(n, m) = (400, 1600)$	$(n, m) = (1600, 6400)$
Projected Gradient	17,788 iterations	487 iterations	7 iterations
Accelerated Gradient	18,412 iterations	506 iterations	7 iterations
Frank-Wolfe	24,137 iterations	333 iterations	26 iterations
Radial Subgradient	8,950,726 iterations	6,835,355 iterations	213,381 iterations
Radial Smoothing	3,827,988 iterations	757,829 iterations	39,005 iterations

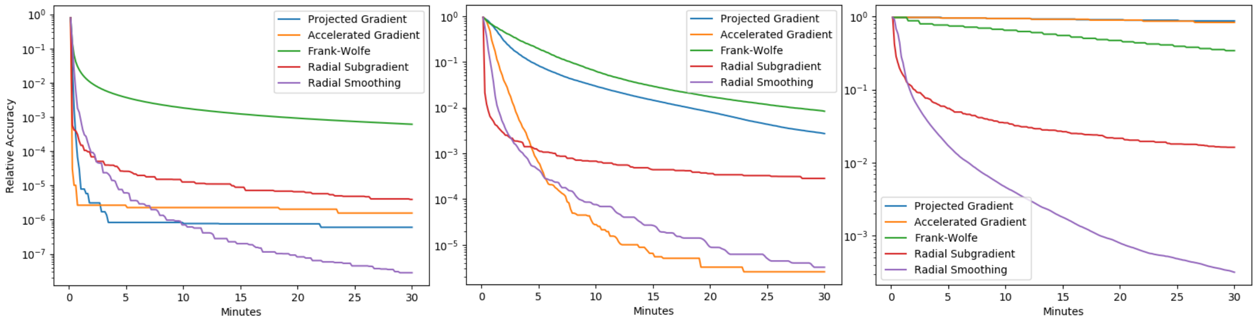
In our largest problem setting $(n, m) = (1600, 6400)$, which has approximately ten million nonzeros, the projected gradient, accelerated gradient, and Frank-Wolfe methods only compute a couple dozen steps within our time budget whereas our radial methods take tens or hundreds of thousands of steps. For any larger problem instances, these classic methods may not even complete a single step and so the radial subgradient and smoothing methods vacuously outperform them.

At every scale of problem size, the radial smoothing method is competitive. For our smallest instance $(n, m) = (100, 400)$, the accelerated and projected gradient methods quickly reach an

[3] Quadratic programming was, in fact, the original motivating setting for the Frank-Wolfe algorithm [13].

[4] There are other QP solvers like OSQP [44] that also only rely on cheap matrix operations, but such operator splitting methods do not maintain a feasible solution at each iteration. Hence they cannot be compared as directly.

[5] The source code is available at github.com/bgrimmer/Radial-Duality-QP-Example



(a) $(n, m) = (100, 400)$ (b) $(n, m) = (400, 1600)$ (c) $(n, m) = (1600, 6400)$

Figure 1: The minimum relative accuracy $\frac{p^* - f(x_k)}{p^*}$ of (6) seen by the projected gradient, accelerated gradient, Frank-Wolfe, radial subgradient and radial smoothing methods over 30 minutes.

accuracy around 10^{-6}, which is the default tolerance of Gurobi, followed shortly afterward by the radial smoothing method. However, for our moderate-sized instance $(n, m) = (400, 1600)$, the classic methods begin to fall off with the radial smoothing method and accelerated method performing comparably. For our largest instance $(n, m) = (1600, 6400)$, the methods relying on orthogonal projection make essentially no progress due to their high iteration cost, and the Frank-Wolfe method only makes minor amounts of progress relative to our radial methods. Our algorithms based on radial duality appear to provide a far more scalable approach.

Throughout our experiments, the radial smoothing method outperforms the radial subgradient method by a couple of orders of magnitude. This agrees with our convergence theory showing that the radial subgradient method converges at a $O(1/\epsilon^2)$ rate while the smoothing technique enables $O(1/\epsilon)$ convergence, presented in Sections 4.1 and 4.2, respectively.

2.2 Broader Computational Advantages of Considering Radially Dual Problems

We conclude this motivating section with a high-level discussion of the computational advantages we see in optimizing over radially dual problem formulations. These benefits all extend beyond the particular radial optimization algorithms considered herein.

2.2.1. Maintaining Primal Feasible Iterates Without Costly Projections Here we generalize the setting of polyhedral constraints considered by (4). After a translation, any convex constraints can be expressed as the intersection of a convex set $S \subseteq \mathcal{E}$ with $0 \in \operatorname{int} S$ and a subspace $T = \{x \in \mathcal{E} \mid Ax = 0\}$. Consider any primal problem with strictly upper radial objective f given by

$$\begin{cases} \max & f(x) \\ \text{s.t.} & x \in S \\ & Ax = 0 \end{cases} = \max_{x \in \mathcal{E}} \min\{f(x), \iota_S(x), \iota_T(x)\}$$

where $\iota_S(x) = \begin{cases} +\infty & \text{if } x \in S \\ 0 & \text{if } x \notin S. \end{cases}$ Then the radially dual problem is

$$\min_{y \in \mathcal{E}} \max\{f^\Gamma(y), \gamma_S(y), \gamma_T(y)\} = \begin{cases} \min & \max\{f^\Gamma(y), \gamma_S(y)\} \\ \text{s.t.} & Ay = 0 \end{cases}$$

where $\gamma_S(y) = \inf\{\lambda \geq 0 \mid y \in \lambda S\}$ denotes the Minkowski gauge since

$$\iota_S^\Gamma(y) = \sup\{v > 0 \mid v \cdot \iota_S(y/v) \leq 1\} = \sup\{v > 0 \mid y/v \notin S\} = \inf\{\lambda > 0 \mid y \in \lambda S\} = \gamma_S(y).$$

Having multiple set constraints $S_1 \dots S_n$ in the primal $\max_{x \in S_1 \cap \dots \cap S_n} f(x)$ simply adds more terms to the radially dual finite maximum of $\min_{y \in \mathcal{E}} \max\{f^\Gamma(y), \gamma_{S_i}(y)\}$.

This formulation allows algorithms to maintain a feasible primal solution at each iteration without requiring costly subproblems relating to S. Instead, a primal feasible solution can be recovered from any radial dual solution $y \in \mathcal{E}$ with $Ay = 0$ as $x = y/\max\{f^\Gamma(y), \gamma_S(y)\} \in S \cap T$. Algorithmically, this replaces the need for orthogonal projections onto the feasible region $S \cap T$ with the cheaper operations of orthogonally projecting onto the subspace T and evaluating the gauge of S. This computational gain was one of the key contributions identified by [38] and was central to the motivation of [39, 16] as well as being a motivation of this work.

2.2.2. Handling Nonconcave Objectives and Nonconvex Constraints Our calculation of the radial dual for quadratic programming did not fundamentally rely on concavity as it also applies to nonconcave problems with a bounded feasible region. Indeed one of the key insights from the first part of this work was divorcing the idea of radial transformations from relying on notions of convexity or concavity. In Section 5.1, we discuss several nonconcave primal maximization problems where radial duality holds, generalizing the above reasoning to star-convex constraints and covering important areas like nonconvex regularization and optimization with outliers.

2.2.3. Efficiently Evaluating Generic Radial Duals A remark on the efficiency of computing the upper radial function transformation $f^\Gamma(y)$: In general, we do not have a closed-form as we found in our quadratic programming example. However, numerically evaluating $f^\Gamma(y)$ is a one-dimensional subproblem that can be solved by bisection whenever f is upper radial (since $v \mapsto vf(y/v)$ is then nondecreasing). Even if f is not upper radial, $f^\Gamma(y)$ may still be tractable to compute. For example, any polynomial f has evaluation of f^Γ amount to polynomial root finding. Once $f^\Gamma(y)$ has been computed, its gradients and Hessian follow from (25) and (26).

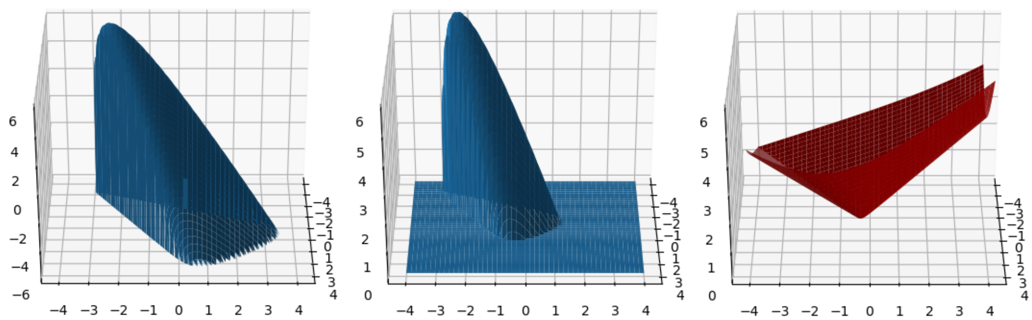
2.2.4. Improving Conditioning and Problem Structure As a final motivating example of the structural advantages of taking the radial dual, consider the following Poisson inverse problem. Given linear measurements with Poisson distribution noise $b_i \sim \texttt{Poisson}(a_i^T x)$, the maximum likelihood estimator is given by maximizing

$$\mathcal{L}(x) := \begin{cases} \sum_i b_i \log(a_i^T x) - a_i^T x & \text{if all } a_i^T x > 0 \\ -\infty & \text{otherwise.} \end{cases}$$

Then given any convex regularizer $r(x)$ and constraint set $S \subseteq \mathbb{R}^n$, we formulate a Poisson inverse problem as

$$\max_{x \in S} \mathcal{L}(x) - r(x). \tag{9}$$

This type of problem arises in image processing (see [4] for a survey of applications from astronomy to medical imaging) as well as in network diffusion and time series modeling (see the many references in [20]). Although this problem is concave, the blow-up from the logarithmic terms prevents standard first-order methods from being applied. Provided the regularization r and constraints S are sufficiently simple, customized primal-dual [20] or Bregman methods [1, 30] provide a powerful tactic for solving this problem.



(a) $\mathcal{L}(x + x_0) - u_0$ (b) $(\mathcal{L}(x + x_0) - u_0)_+$ (c) $(\mathcal{L}(x + x_0) - u_0)_+^\Gamma(y)$

Figure 2: Example (a) translating, (b) truncating, and then (c) taking the radial dual of (9).

For generic S and r, our radial duality can be applied. Given any $x_0 \in \operatorname{int} \operatorname{dom} \mathcal{L} \cap S$ and $u_0 < \mathcal{L}(x_0) - r(x_0)$, we can reformulate this objective function to be strictly upper radial via a simple translation and truncation. We consider the equivalent problem of

$$\max_{x \in \mathbb{R}^n} \min\{(\mathcal{L}(x + x_0) - r(x + x_0) - u_0)_+, \iota_S(x + x_0)\}.$$

Then we can employ our radial duality machinery using [17, Proposition 3.8] since our translated and truncated objective is concave with 0 strictly in its domain. The radial dual here is defined everywhere $\operatorname{dom} f^\Gamma = \mathbb{R}^n$, is globally uniformly Lipschitz continuous (see Proposition 3.1) and if $S = \mathbb{R}^n$ and $r(x)$ is twice continuously differentiable, has globally Lipschitz continuous gradient (see Corollary 3.3). The primal formulation is none of these. Note that different translations of the objective (here corresponding to a different choice of (x_0, u_0)) produce different radial duals, which in turn can have very different global Lipschitz and smoothness constants.

Figure 2 shows the steps of taking the radial dual of a two-dimensional likelihood maximization problem with $\{a_1, a_2, a_3\} = \{(2, -1), (1, 1), (-1, 2)\}$, $b_i = 1$, $S = \mathbb{R}^2$, and $r = 0$: (a) shows the translated objective with $x_0 = (3, 3)$ and $u_0 = -10$, (b) shows the truncated strictly upper radial nonnegative optimization problem, and (c) shows the well behaved radially dual objective.

2.3 Notation and Review

We consider functions $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, where $\overline{\mathbb{R}}_{++} = \mathbb{R}_{++} \cup \{0, +\infty\}$ denotes the "extended positive reals". Here 0 and $+\infty$ are the limit objects of $\mathbb{R}_{++}$, mirroring the roles of $-\infty$ and $+\infty$ in the extended reals. The effective domain, graph, epigraph, and hypograph of such a function are

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

We say a function $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is upper (lower) semicontinuous if hypo f (epi f) is closed with respect to $\mathcal{E} \times \mathbb{R}_{++}$. Equivalently, a function is upper semicontinuous if for all $x \in \mathcal{E}$, $f(x) = \limsup_{x' \to x} f(x')$ and lower semicontinuous if $f(x) = \liminf_{x' \to x} f(x')$. We say a function $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$ is concave (convex) if hypo f (epi f) is convex. The set of *convex normal vectors* of a set

$S \subseteq \mathcal{E} \times \mathbb{R}$ at some $(x,u) \in S$ is denoted by

$$N_S^C((x,u)) := \{(\zeta,\delta) \mid (\zeta,\delta)^T((x,u)-(x',u')) \geq 0 \ \forall (x',u') \in S\}.$$

Then the *convex subdifferential* and *convex supdifferential* of a function f is denoted by

$$\partial_C f(x) := \{\zeta \mid (\zeta,-1) \in N^C_{\text{epi } f}((x,f(x)))\},$$
$$\partial^C f(x) := \{\zeta \mid (-\zeta,1) \in N^C_{\text{hypo } f}((x,f(x)))\}.$$

We also consider the generalization given by proximal normal vectors and sub/supdifferentials of

$$N_S^P((x,u)) := \{(\zeta,\delta) \mid (x,u) \in \text{proj}_S((x,u)+\epsilon(\zeta,\delta)) \text{ for some } \epsilon > 0\},$$
$$\partial_P f(x) := \{\zeta \mid (\zeta,-1) \in N^P_{\text{epi } f}((x,f(x)))\},$$
$$\partial^P f(x) := \{\zeta \mid (-\zeta,1) \in N^P_{\text{hypo } f}((x,f(x)))\}.$$

Dual Families of Functions. Most of our theory characterizing the radial transformation relies on the given function being (strictly) upper radial. Recall that [17, Proposition 3.3] shows an upper semicontinuous function f is upper radial (that is, our radial duality $f^{\Gamma\Gamma} = f$ holds) if and only if all $(x,u) \in \text{hypo } f$ and $(\zeta,\delta) \in N^P_{\text{hypo } f}((x,u))$ satisfy

$$(\zeta,\delta)^T(x,u) \geq 0. \tag{10}$$

Geometrically, this corresponds to the origin lying below all of the hyperplanes induced by proximal normal vectors of the hypograph. Similarly, [17, Proposition 3.5] ensures a continuously differentiable function f is strictly upper radial if all $x \in \text{dom } f$ satisfy

$$(\nabla f(x),-1)^T(x,u) < 0. \tag{11}$$

For concave functions, being upper radial corresponds to the origin lying in the function's domain. In particular, [17, Proposition 3.8] ensures an upper semicontinuous concave function f is strictly upper radial if

$$0 \in \text{int } \{x \mid f(x) > 0\}. \tag{12}$$

Assuming strict upper radiality holds, the following families of functions are radially dual

$$f \text{ is upper semicontinuous} \iff f^\Gamma \text{ is lower semicontinuous,} \tag{13}$$
$$f \text{ is continuous} \iff f^\Gamma \text{ is continuous,} \tag{14}$$
$$f \text{ is concave} \iff f^\Gamma \text{ is convex,} \tag{15}$$

where these follow from [17, Propositions 3.13, 3.15]. For differentiable functions satisfying (11), [17, Proposition 4.5] shows

$$f \text{ is } k \text{ times differentiable} \iff f^\Gamma \text{ is } k \text{ times differentiable,} \tag{16}$$
$$f \text{ is analytic} \iff f^\Gamma \text{ is analytic.} \tag{17}$$

Relating Extreme Points, (Sub)Gradients, and Hessians. We recall a few bijections relating functions and their radial transformations. For any strictly upper radial f, [17, Lemma 4.1] ensures

$$\text{epi } f^\Gamma = \Gamma(\text{hypo } f). \tag{18}$$

Further, [17, Lemma 4.4] shows for any continuous strictly upper radial function, the following pair of bijections between graphs and domains hold

$$\text{graph } f^\Gamma = \Gamma(\text{graph } f), \tag{19}$$

$$y \in \text{dom } f^\Gamma \iff y/f^\Gamma(y) \in \text{dom } f. \tag{20}$$

Then [17, Propositions 4.9, 4.10] shows that the radial point transformation relates the maximizers of a strictly upper radial function f to the minimizers of f^Γ as well as relates their stationary points

$$\text{argmin } f^\Gamma \times \{\inf f^\Gamma\} = \Gamma(\text{argmax } f \times \{\sup f\}), \tag{21}$$

$$\{(y, f^\Gamma(y)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial_P f^\Gamma(y)\} = \Gamma\{(x, f(x)) \in \mathcal{E} \times \mathbb{R}_{++} \mid 0 \in \partial^P f(x)\}. \tag{22}$$

In particular, for any upper semicontinuous, strictly upper radial f, the convex and proximal subgradients of its upper radial transformation are given by [17, Propositions 4.2, 4.3] as

$$\partial_C f^\Gamma(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix}\zeta\\ \delta\end{bmatrix} \in N^C_{\text{hypo } f}((x,u)),\ (\zeta,\delta)^T(x,u) > 0 \right\} \tag{23}$$

$$\partial_P f^\Gamma(y) = \left\{ \frac{\zeta}{(\zeta,\delta)^T(x,u)} \mid \begin{bmatrix}\zeta\\ \delta\end{bmatrix} \in N^P_{\text{hypo } f}((x,u)),\ (\zeta,\delta)^T(x,u) > 0 \right\} \tag{24}$$

where $(x, u) = \Gamma(y, f^\Gamma(y))$. Further, if f is continuously differentiable and satisfies (11), [17, Proposition 4.5] shows the gradient of the upper radial transformation at $y = x/f(x)$ is

$$\nabla f^\Gamma(y) = \frac{\nabla f(x)}{(\nabla f(x), -1)^T(x, f(x))}. \tag{25}$$

If in addition we suppose f is twice continuously differentiable around x, [17, Proposition 4.6] shows the Hessian of the upper radial transformation is

$$\nabla^2 f^\Gamma(y) = \frac{f(x)}{(\nabla f(x), -1)^T(x, f(x))} \cdot J\nabla^2 f(x) J^T \tag{26}$$

where $J = I - \frac{\nabla f(x) x^T}{(\nabla f(x),-1)^T(x,f(x))}$.

3 Conditioning of the Radially Dual Problem

As we have seen, the radial dual often enjoys very favorable structural properties. In the following three subsections, we characterize the Lipschitz continuity, smoothness, and growth conditions of the radially dual problem. Historically, these properties are all of great importance to the development of first-order optimization algorithms.

3.1 Lipschitz Continuity of the Radially Dual Problem

We say a function f is uniformly M-Lipschitz continuous if for all $x, x' \in \mathcal{E}$,

$$|f(x) - f(x')| \le M\|x - x'\|.$$

For any lower semicontinuous function $f\colon \mathcal{E} \to \mathbb{R}_{++} \cup \{\infty\}$, M-Lipschitz continuity is equivalent to all proximal subgradients $\zeta \in \partial_P f(x)$ having norm bounded by M [9, Theorem 1.7.3].

Lipschitz continuity plays an important role in the analysis of many first-order methods for nonsmooth optimization. Recalling that the previous works [38, 39, 16] critically rely on their radially reformulated objective being uniformly Lipschitz, here we present a general characterization of when the radial transformation of a function is uniformly Lipschitz. To take advantage of the second characterization of Lipschitz continuity above, we need to ensure f^Γ maps into $\mathbb{R}_{++} \cup \{\infty\}$. The following simple assumption is equivalent to this (by the definition of the upper radial transformation): for all $y \in \mathcal{E}$

$$\lim_{v \to 0} v \cdot f(y/v) = 0.$$

This condition is always the case when f is bounded above as will typically be the case for our primal maximization problem. Under this condition, we find that the Lipschitz continuity of f^Γ is controlled by the distance (measured in $\mathcal{E}$) from the origin to each hyperplane defined by a proximal normal vector:

$$R(f) = \inf\{\|x'\| \mid (\zeta, \delta) \in N^P_{\text{hypo } f}(x, u),\ (\zeta, \delta)^T((x', 0) - (x, u)) = 0\}.$$

The following proposition gives the exact Lipschitz constant in terms of $R(f)$.

Proposition 3.1. *Consider any upper semicontinuous, strictly upper radial f where all $y \in \mathcal{E}$ have $\lim_{v\to 0} v \cdot f(y/v) = 0$. Then f^Γ is $1/R(f)$-Lipschitz continuous.*

Proof. The key observation here is that for any $(x, u) \in \text{hypo } f$ and $(\zeta, \delta) \in N^P_{\text{hypo } f}((x, u))$,

$$\begin{aligned}(\zeta, \delta)^T(x, u) &= \inf\big\{\zeta^T x' \mid (\zeta, \delta)^T((x', 0) - (x, u)) = 0\big\} \\ &= \|\zeta\| \inf\big\{\|x'\| \mid (\zeta, \delta)^T((x', 0) - (x, u)) = 0\big\} \\ &\geq \|\zeta\| R(f)\end{aligned}$$

where the first equality is trivial and the second uses that the minimum norm point in this hyperplane will be a multiple of ζ. Then the subgradient formula (24) ensures any $\zeta' \in \partial_P f^\Gamma(y)$ must have

$$\|\zeta'\| = \frac{\|\zeta\|}{(\zeta, \delta)^T(x, u)} \leq 1/R(f)$$

for $(x, u) = \Gamma(y, f^\Gamma(y))$ and some $(\zeta, \delta) \in N^P_{\text{hypo } f}((x, u))$. Since every radially dual subgradient is uniformly bounded, f^Γ is uniformly Lipschitz. Considering a sequence of $(\zeta, \delta) \in N^P_{\text{hypo } f}((x, u))$ approaching attainment of $R(f)$ makes this argument tight. □

The condition $(x, u)^T(\zeta, \delta) \geq R(f)\|\zeta\|$ can be viewed as a natural way to quantify how radial f is by strengthening (10). When f is concave, $R(f)$ can be simplified to

$$R(f) = \inf\{\|x\| \mid f(x) = 0\}, \tag{27}$$

which matches the Lipschitz constants used in the previous works [38, 39, 16]. This gives a natural way to measure the extent of radiality of a concave function by strengthening (12). From this, we see any concave maximization problem (with a known point in the interior of its domain) can be translated and transformed into a convex minimization problem that is uniformly Lipschitz continuous with constant depending on how interior the known point is to the function's domain.

3.2 Smoothness of the Radially Dual Problem

We say a continuously differentiable function f is uniformly L-smooth if its gradient is L-Lipschitz continuous: for all $x, x' \in \text{dom } f$

$$\|\nabla f(x) - \nabla f(x')\| \leq L\|x - x'\|.$$

As an example, consider the radial dual of the continuously differentiable function $f(x) = \sqrt{1 - x^T Q x}_+$, which is upper radial for any matrix Q. This radially transforms into the similarly shaped function

$$f^\Gamma(y) = \sup\{v > 0 \mid v\sqrt{1 - y^T Q y/v^2} \leq 1\} = \sup\{v > 0 \mid v^2 - y^T Q y \leq 1\} = \sqrt{1 + y^T Q y}_+.$$

Supposing Q is positive semidefinite and nonzero, our primal is concave and differentiable on its domain but fails to have a Lipschitz gradient since $\nabla f(x)$ blows up at the boundary of its domain. However, in this case, the radially dual f^Γ is well behaved, being convex and $\lambda_{max}(Q)$-smooth.

For generic functions, we cannot hope to find smoothness out of thin air (like we do in the above example or quite generically with Lipschitz continuity in the previous section). This is due to (16) which establishes differentiability is preserved under the radial transformation. In line with this equivalence, we find that when f is L-smooth, f^Γ is $O(L)$-smooth, provided the domain of f is bounded. Let $D(f) = \sup\{\|x\| \mid x \in \text{dom } f\}$ denote the norm of the largest point in the domain of f. Note that since we are primarily taking the radial dual of maximization problems that are bounded above and truncated below to be nonnegative optimization, $D(f)$ can be viewed as bounding the level set dom $f = \{x \mid f(x) > 0\}$.

The following proposition shows the operator norm of the radial transformation's Hessian is controlled by the ratio between $D(f)$ and $R(f)$ and the norm of the primal Hessian. From this, we conclude for twice differentiable L-smooth functions, the radial dual is also $O(L)$-smooth.

Proposition 3.2. *Consider any upper radial f with $D(f) < \infty$ and $R(f) > 0$ and $x, y \in \mathcal{E}$ satisfying $(x, f(x)) = \Gamma(y, f^\Gamma(y))$. If f is twice continuously differentiable around x, then*

$$\|\nabla^2 f^\Gamma(y)\| \leq \left(1 + \frac{D(f)}{R(f)}\right)^3 \|\nabla^2 f(x)\|.$$

Proof. First we verify that $(\nabla f(x), -1)^T(x, f(x)) < 0$ holds for all $x \in \text{dom } f$ and so the Hessian formula (26) can be applied: if $\nabla f(x) = 0$, $(\nabla f(x), -1)^T(x, f(x)) = -f(x) < 0$ and if $\nabla f(x) \neq 0$, $(\nabla f(x), -1)^T(x, f(x)) \leq -\|\nabla f(x)\| R(f) < 0$. Then our bound on the Hessian of f^Γ follows from the following pair of inequalities. First, we have

$$\frac{f(x)}{(\nabla f(x), -1)^T(x, f(x))} = 1 + \frac{\nabla f(x)^T x}{(\nabla f(x), -1)^T(x, f(x))} \leq 1 + \frac{\|\nabla f(x)\|\|x\|}{|(\nabla f(x), -1)^T(x, f(x))|} \leq 1 + \frac{\|x\|}{R(f)}.$$

Second, the matrix $J = I - \frac{\nabla f(x) x^T}{(\nabla f(x), -1)^T(x, f(x))}$ has operator norm bounded by

$$\|J\| \leq 1 + \frac{\|\nabla f(x)\|\|x\|}{|(\nabla f(x), -1)^T(x, f(x))|} \leq 1 + \frac{\|x\|}{R(f)}.$$

Then applying our bounds to each term in the Hessian formula (26) gives the claimed result. □

Corollary 3.3. *Consider any upper radial, twice continuously differentiable f with $D(f) < \infty$ and $R(f) > 0$. If f is L-smooth, then f^Γ is $\left(1 + \frac{D(f)}{R(f)}\right)^3 L$-smooth.*

Proof. For a twice continuously differentiable function, having L-Lipschitz gradient is equivalent to having Hessian bounded in operator norm by L. Noting that $R(f) > 0$ implies f is strictly upper radial by (11), we have a bijection between the domains of f and f^Γ from (20). Hence the Hessian of f^Γ is uniformly bounded by $\left(1 + \frac{D(f)}{R(f)}\right)^3 L$. $\square$

Although this result requires smoothness of the primal objective f to be maximized, it still provides an algorithmically valuable tool due to the symmetry-breaking nature of considering functions on the extended positive reals $\overline{\mathbb{R}}_{++}$. Supposing f is bounded above, this result allows us to extend smoothness of f on a level set dom $f = \{x \mid f(x) > 0\}$ to global smoothness of the dual f^Γ on dom $f^\Gamma = \mathcal{E}$.

For example, consider an unconstrained $S = \mathbb{R}^n$ instance of our previous motivating example of the Poisson likelihood problem (9) which is not defined everywhere (only on $\{x \mid a_i^T x > 0\}$) with gradients blowing up as x approaches the boundary of this domain. However, provided the measurements $\{a_i\}$ span $\mathbb{R}^n$, this objective has bounded level sets. Consequently, for any twice continuously differentiable $r(x)$, our radial duality provides a globally smooth reformulation.

3.3 Growth Conditions in the Radially Dual Problem

For a lower semicontinuous function $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, we say the Łojasiewicz condition holds at a local minimum x^* if for some constants $r > 0$, $C > 0$ and exponent $\theta \in [0, 1)$, all nearby $x \in B(x^*, r)$ have

$$\mathrm{dist}(0, \partial_P f(x)) \geq C(f(x) - f(x^*))^\theta. \tag{28}$$

For an upper semicontinuous function f with local maximum x^*, we instead require all nearby $x \in B(x^*, r)$ have

$$\mathrm{dist}(0, \partial^P f(x)) \geq C(f(x^*) - f(x))^\theta. \tag{29}$$

These conditions are widespread, holding for generic subanalytic functions [29, 28] and nonsmooth subanalytic convex functions [5]. These properties are closely related to the Kurdyka-Łojasiewicz condition [24] and Hölderian growth/error bounds used by [6, 46, 41, 40], which are known to speed up the convergence of many first-order methods.

Under mild conditions, the Łojasiewicz condition is preserved by our radial transformation. Consequently, optimization algorithms based on solving the radially dual problem can enjoy the same improved convergence historically expected in the primal from such conditions.

Proposition 3.4. *Consider any upper semicontinuous, strictly upper radial function f with $R(f) > 0$ and $\sup f \in \mathbb{R}_{++}$. If f satisfies the Łojasiewicz condition* (29) *at some $x^* \in \mathrm{dom}\ f$ with exponent θ, then f^Γ at $y^* = x^*/f(x^*) \in \mathrm{dom}\ f^\Gamma$ satisfies the Łojasiewicz condition* (28) *with the same exponent θ.*

Proof. Let r, C, θ satisfy the Łojasiewicz condition of f at x^* and denote the radially dual point as $y^* = x^*/f(x^*)$. Since f is bounded above, f^Γ is $1/R(f)$-Lipschitz continuous by Proposition 3.1. Then every $0 < r' < f^\Gamma(y^*)R(f)$ has $y \in B(y^*, r')$ map to $x = y/f^\Gamma(y)$ with

$$\begin{aligned}
\|x - x^*\| &= \left\| \frac{y}{f^\Gamma(y)} - \frac{y^*}{f^\Gamma(y^*)} \right\| \\
&\leq \left\| \frac{y - y*}{f^\Gamma(y)} \right\| + \left\| \frac{y^*}{f^\Gamma(y)} - \frac{y^*}{f^\Gamma(y^*)} \right\| \\
&= \frac{\|y - y^*\|}{f^\Gamma(y)} + \frac{\|y^*\| |f^\Gamma(y) - f^\Gamma(y^*)|}{f^\Gamma(y) f^\Gamma(y^*)}
\end{aligned}$$

$$\le \frac{r'}{f^\Gamma(y)} + \frac{\|y^*\|r'}{R(f)f^\Gamma(y)f^\Gamma(y^*)}$$
$$\le \frac{r'}{f^\Gamma(y^*) - r'/R(f)} + \frac{\|y^*\|r'}{R(f)(f^\Gamma(y^*) - r'/R(f))f^\Gamma(y^*)}.$$

Therefore selecting small enough r' guarantees that all of the dual points near y^* map back to primal points $x = y/f^\Gamma(y)$ in the ball $B(x^*, r)$ where the Łojasiewicz condition holds. Further the Lipschitz continuity of the radial dual allows us to guarantee that all of these primal points have $f(x)$ bounded below by nearly $f(x^*)$ as

$$f(x) = f^{\Gamma\Gamma}(x) \ge 1/f^\Gamma(y) \ge 1/(f^\Gamma(y^*) - r'/R(f)) = (f(x^*)^{-1} + (R(f)/r')^{-1})^{-1}.$$

Combining this with the assumed upper semincontinuity of f, we have $f(x) \to f(x^*)$ as $y \to y^*$ (despite not assuming continuity of the primal function f).

Then all that remains is to show the Łojasiewicz supgradient norm lower bound from the primal extends to lower bound the norm of the radially dual subgradients. For every $y \in B(y^*, r')$, the formula (24) ensures every $\zeta' \in \partial_P f^\Gamma(y)$ has $\zeta' = \zeta/(\zeta,\delta)^T(x,u)$ where $(x,u) = \Gamma(y, f^\Gamma(y))$ and $(\zeta,\delta) \in N^P_{\text{hypo } f}((x,u))$. First, suppose $\delta \ne 0$. Then $u = f(x)$ and $-\zeta/\delta \in \partial^P f(x)$ is a primal supgradient. Consequently, we can bound the size of our radially dual subgradient as

$$\begin{aligned}
\|\zeta'\| &= \frac{\|\zeta/\delta\|}{(\zeta/\delta, 1)^T(x, f(x))} \\
&\ge \frac{\|\zeta/\delta\|}{\|\zeta/\delta\|\|x\| + f(x)} \\
&\ge \frac{C(f(x^*) - f(x))^\theta}{C(f(x^*) - f(x))^\theta\|x\| + f(x)} \\
&= \frac{Cf^\theta(x)f^\theta(x^*)}{C(f(x^*) - f(x))^\theta\|x\| + f(x)}\left(f^\Gamma(y) - f^\Gamma(y^*)\right)^\theta
\end{aligned}$$

where the final inequality uses that $f(x) \ge 1/f^\Gamma(y)$ and $f(x^*) = 1/f^\Gamma(y^*)$. Recalling that as $y \to y^*$, the related primal point $x = y/f^\Gamma(y) \to x^*$ and $f(x) \to f(x^*)$, the coefficient above must converge to a positive constant

$$\frac{Cf^\theta(x)f^\theta(x^*)}{C(f(x^*) - f(x))^\theta\|x\| + f(x)} \to \frac{Cf^{2\theta}(x^*)}{C0^\theta\|x^*\| + f(x^*)}.$$

The boundary case of horizontal normal vectors with $\delta = 0$ follows from the same argument above by passing to a sequence of points $(x_i, f(x_i)) \to (x, f(x))$ and proximal normal vectors $(\zeta_i, \delta_i) \in N^P_{\text{hypo } f}((x_i, f(x_i)))$ with $(\zeta_i, \delta_i) \to (\zeta, \delta)$ and $\delta_i \ne 0$. The existence of such a sequence is guaranteed by the Horizontal Approximation Theorem [9, Page 67]. □

The case of $\theta = 0$ above is an important special case known as sharpness. If this condition holds globally, (28) and (29) correspond to the following global error bounds holding for all $x \in \mathcal{E}$

$$f(x) \ge f(x^*) + C\|x - x^*\| \tag{30}$$

and

$$f(x) \le (f(x^*) - C\|x - x^*\|)_+ \tag{31}$$

respectively. This condition has a long history in nonsmooth optimization (see Burke and Ferris [7] as a classic reference establishing the prevalence of sharp minima). The two global sharp error bounds (31) and (30) are dually related as follows.

Proposition 3.5. *For any upper semicontinuous, strictly upper radial f satisfying* (31) *at $x^* \in \mathcal{E}$ with constant C, then f^Γ satisfies* (30) *at $y^* = x^*/f(x^*)$ with constant*

$$\frac{C}{C\|x^*\| + f(x^*)}.$$

Proof. Denote the assumed upper bound on f from sharpness as $h(x) := f(x^*) - C\|x - x^*\|$. Then h_+ must be strictly upper radial due to (12) since h is concave with $h(0) > 0$ as

$$h(0) = 2h(x^*/2) - h(x^*) \geq 2f(x^*/2) - f(x^*) > 0$$

where the first equality uses that h is linear on the segment $[0, x^*]$, the inequality uses that $h(x^*/2) \geq f(x^*/2)$, and the strict inequality uses that f is strictly upper radial. The upper radial transformation h_+^Γ is lower bounded by our claimed sharpness lower bound for any $y \in \mathcal{E}$

$$h_+^\Gamma(y) \geq \frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|} = f^\Gamma(y^*) + \frac{C\|y - y^*\|}{f(x^*) + C\|x^*\|}$$

since $h_+^p(y, v)$ at $v = \frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*)+C\|x^*\|}$ has

$$\begin{aligned}
h_+^p(y, v) &= \left(\frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|}\right) f(x^*) - C\left\| y - \left(\frac{1}{f(x^*)} + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|}\right) x^* \right\| \\
&\leq 1 + \frac{C\|y - x^*/f(x^*)\|}{f(x^*) + C\|x^*\|} f(x^*) - C\|y - x^*/f(x^*)\| + \frac{C^2\|y - x^*/f(x^*)\|\|x^*\|}{f(x^*) + C\|x^*\|} \\
&= 1 + C\|y - x^*/f(x^*)\| \left(\frac{f(x^*)}{f(x^*) + C\|x^*\|} - 1 + \frac{C\|x^*\|}{f(x^*) + C\|x^*\|}\right) \\
&= 1
\end{aligned}$$

where the single inequality above uses the reverse triangle inequality. Using [17, Lemma 4.7], $f \leq h_+$ implies $f^\Gamma \geq h_+^\Gamma$, competing our proof. ☐

4 Radial Algorithms for Concave Maximization

Now we turn our attention to understanding the primal convergence guarantees that follow from algorithms minimizing the radial dual. In this section, we consider concave maximization problems where being strictly upper radial and having $R(f) > 0$ hold without loss of generality via a simple translation. Then the following section tackles nonconcave maximization problems where more care must be taken to ensure our duality holds.

We first remark on the natural measure of optimality in the primal that arise from considering the radial dual. Recall the set of fixed points of Γ are exactly the horizontal line at height one $\{(y, 1) \mid x \in \mathcal{E}\} = \Gamma\{(x, 1) \mid x \in \mathcal{E}\}$. Consequently, a natural way to relate nearly optimal solutions between then primal and radial dual comes from considering when $\sup f = \inf f^\Gamma = 1$. In this case, finding a dual point with accuracy

$$f^\Gamma(y_k) - \inf f^\Gamma \leq \epsilon$$

is equivalent to the relative accuracy primal guarantee of

$$\frac{\sup f - f(x_k)}{f(x_k)} \leq \epsilon.$$

using that $1/f^\Gamma(y_k) \le f^{\Gamma\Gamma}(x_k) = f(x_k)$ for $x_k = y_k/f^\Gamma(y_k)$ on any upper radial f. Following from this, we state all of our radial algorithm convergence guarantees in relative terms.

Secondly, we remark on the meaning of finding a radially dual solution minimized all the way to zero objective value $f^\Gamma(y) = 0$. In this case, y certifies that the primal maximization is unbounded as the ray $(y,1)/v \in \text{epi } f$ for all $v > 0$. Note the converse of this is not true: for example, the strictly radial function $f(x) = \sqrt{x+1}_+$ is unbounded above, but has $f^\Gamma(y) > 0$ everywhere.

4.1 Radial Subgradient Method

We begin by considering the radial subgradient method previously defined in Algorithm 1. This method simply takes the radial dual, applies the classic subgradient method to the resulting minimization problem, and then takes the radial dual again to return a primal solution. Importantly this method is projection-free since any primal constraint set S appears in the radial dual objective through its gauge γ_S. This method is very similar to those considered in [38, 16] which also apply a subgradient method to a radial reformulation. However, those methods include additional steps periodically rescaling their radial objective. Our algorithm omits such steps while matching the improved convergence guarantees of [16].

The standard subgradient method analysis shows the radial subgradient iterates y_k converge in terms of radial dual optimality at a rate controlled by the radially dual Lipschitz constant. Recall that translating a point in the interior of hypo f to the origin ensures $R(f) > 0$ and so the radial dual is Lipschitz continuous by Proposition 3.1). Consequently, no structure needs to be assumed beyond concavity to analyze the radial subgradient method.

Theorem 4.1. *Consider any upper semicontinuous, concave f with $R(f) > 0$ and $p^* = \sup f \in \mathbb{R}_{++}$ attained on some nonempty set $X^* \subseteq \mathcal{E}$. Then the radial subgradient method (Algorithm 1) with stepsizes α_k has primal solutions $x_k = y_k/f^\Gamma(y_k)$ satisfy*

$$\min_{k<T}\left\{\frac{p^* - f(x_k)}{f(x_k)}\right\} \le \frac{\text{dist}(p^* y_0, X^*)^2 + \sum_{k=0}^{T-1}(p^*\alpha_k/R(f))^2}{2\sum_{k=0}^{T-1} p^*\alpha_k}.$$

Selecting $x_0 = 0$ and $\alpha_k = \epsilon f^\Gamma(y_k)/\|\zeta_k'\|^2$ for any $\epsilon > 0$ ensures

$$T \ge \frac{\text{dist}(x_0, X^*)^2}{R(f)^2\epsilon^2} \quad \Longrightarrow \quad \frac{1}{T}\sum_{k=0}^{T-1}\frac{p^* - f(x_k)}{p^*} \le \epsilon.$$

Proof. Having $R(f) > 0$ ensures f is strictly upper radial by (12). Then f^Γ is convex by (15) and has minimum value $d^* = 1/p^*$ attained on $Y^* := X^*/p^*$ by (21). The classic convex convergence analysis of subgradient methods follows from the fact that: for any $y^* \in Y^*$,

$$\begin{aligned}\|y_{k+1} - y^*\|^2 &= \|y_k - y^*\|^2 - 2\alpha_k \zeta_k'^T(y_k - y^*) + \alpha_k^2\|\zeta_k'\|^2 \\ &\le \|y_k - y^*\|^2 - 2\alpha_k(f^\Gamma(y_k) - d^*) + \alpha_k^2\|\zeta_k'\|^2\end{aligned}$$

and so inductively,

$$\sum_{k=0}^{T-1}\alpha_k(f^\Gamma(y_k) - d^*) \le \frac{\|y_0 - y^*\|^2 + \sum_{k=0}^{T-1}\alpha_k^2\|\zeta_k'\|^2}{2}. \tag{32}$$

Noting $(x_k, u_k) = \Gamma(y_k, f^\Gamma(y_k))$, the primal iterates have $f(x_k) \ge 1/f^\Gamma(y_k)$. Then multiplying through by $(1/d^*)^2$, which equals $(p^*)^2$, we arrive at the relative primal convergence rate of

$$\sum_{k=0}^{T-1}\frac{\alpha_k}{d^*}\left(\frac{p^* - f(x_k)}{f(x_k)}\right) = \sum_{k=0}^{T-1}\frac{\alpha_k}{(d^*)^2}\left(\frac{1}{f(x_k)} - \frac{1}{p^*}\right) \le \frac{\|y_0/d^* - y^*/d^*\|^2 + \sum_{k=0}^{T-1}(\alpha_k/d^*)^2\|\zeta_k'\|^2}{2}.$$

Since f^Γ is $1/R(f)$-Lipschitz (by Proposition 3.1), every radially dual subgradient is uniformly bounded by $\|\zeta'_k\| \leq 1/R(f)$. Then selecting $y^* = \mathrm{proj}_{Y^*}(y_0)$ gives our claimed primal convergence rate. Observe that setting $x_0 = 0$ sets $y_0 = x_0/f(x_0) = 0$ as well. Then plugging $\alpha_k = \epsilon f^\Gamma(y_k)/\|\zeta'_k\|^2$ into (32) yields

$$\begin{aligned}\frac{\mathrm{dist}(x_0, X^*)^2}{2} = \frac{\mathrm{dist}(y_0/d^*, X^*)^2}{2} &\geq \sum_{k=0}^{T-1} \frac{\alpha_k}{d^*}\left(\frac{f^\Gamma(y_k) - d^*}{d^*} - \frac{1}{2}\left(\frac{\alpha_k}{d^*}\right)\|\zeta'_k\|^2\right) \\ &\geq \sum_{k=0}^{T-1} \epsilon\left(\frac{f^\Gamma(y_k)}{d^*\|\zeta'_k\|}\right)^2\left(\frac{p^* - f(x_k)}{p^*} - \frac{\epsilon}{2}\right) \\ &\geq \sum_{k=0}^{T-1} \epsilon R(f)^2\left(\frac{p^* - f(x_k)}{p^*} - \frac{\epsilon}{2}\right).\end{aligned}$$

Rearranging this completes our proof as the primal convergence guarantee becomes

$$\frac{1}{T}\sum_{k=0}^{T-1} \frac{p^* - f(x_k)}{p^*} \leq \frac{\mathrm{dist}(x_0, X^*)^2}{2R(f)^2\epsilon T} + \frac{\epsilon}{2}. \qquad \square$$

Recall for concave f the formula for $R(f)$ can be simplified to $\inf\{\|x\| \mid f(x) = 0\}$, which quantifies how interior the origin is to the set $\{x \mid f(x) > 0\}$. In this light, the constants in this rate agree with those in the guarantees of [16], up to small constants.

The classic convergence rates of the subgradient method improve in the presence of growth conditions like (28) or (30). For example growth with exponent $\theta = 1/2$ corresponds to the case of quadratic growth (generalizing strong convexity) and leads to faster $O(1/\epsilon)$ convergence, see [25] as a simple example. When $\theta = 0$, sharp growth enables the classic subgradient method to converge linearly, as shown by Polyak [36, 37] more than 50 years ago. Recalling that these quantities are preserved from primal to radial dual (Propositions 3.4 and 3.5), we find the same improvements to hold for our radial subgradient method. The following two theorems establish this speed up when $\theta = 0$ and $\theta > 0$, using the radially dual Polyak stepsize $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta'_k\|^2$.

Theorem 4.2. *Consider any upper semicontinuous, concave f with $R(f) > 0$ and $p^* = \sup f \in \mathbb{R}_{++}$ attained at $x^* \in \mathcal{E}$. Fixing $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta'_k\|^2$, if f satisfies the sharp growth condition* (31)*, then the radial subgradient method (Algorithm 1) has $x_k = y_k/f^\Gamma(y_k)$ satisfy*

$$T \geq 4\left(\frac{p^* + C\|x^*\|}{CR(f)}\right)^2 \log_2\left(\frac{p^* - f(x_0)}{f(x_0)\epsilon}\right) \implies \min_{k<T}\left\{\frac{p^* - f(x_k)}{f(x_k)}\right\} \leq \epsilon.$$

Proof. Plugging the stepsize choice $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta'_k\|^2$ into (32) implies

$$\sum_{k=0}^{T-1} \frac{(f^\Gamma(y_k) - d^*)^2}{2} \leq \frac{\|y_0 - y^*\|^2}{2R(f)^2} \tag{33}$$

where $y^* = x^*/p^*$ and Proposition 3.1 is used to bound $\|\zeta'_k\| \leq 1/R(f)$. Then the radially dual sharpness bound from Proposition 3.5 guarantees $\|y_0 - y^*\| \leq \frac{p^*+C\|x^*\|}{C}(f^\Gamma(y_0) - d^*)$. Hence

$$\frac{1}{T}\sum_{k=0}^{T-1}(f^\Gamma(y_k) - d^*)^2 \leq \frac{(p^* + C\|x^*\|)^2(f^\Gamma(y_0) - d^*)^2}{C^2R(f)^2T}.$$

Therefore some $k \leq 4\left(\frac{p^*+C\|x^*\|}{CR(f)}\right)^2$ has halved the dual objective gap, $f^\Gamma(y_k)-d^* \leq (f^\Gamma(y_0)-d^*)/2$. Repeatedly applying this, we conclude that for any $\epsilon' > 0$,

$$T \geq 4\left(\frac{p^* + C\|x^*\|}{CR(f)}\right)^2 \log_2\left(\frac{f^\Gamma(y_0) - d^*}{\epsilon'}\right) \implies \min_{k<T}\left\{f^\Gamma(y_k) - d^*\right\} \leq \epsilon'.$$

Considering $\epsilon' = \epsilon/p^*$ gives the claimed linear convergence rate. □

This generalizes the linear convergence results shown by [38] for linear programming. To the best of our knowledge, this is the first first-order method linear convergence guarantee for generic non-Lipschitz, sharp convex optimization.

Theorem 4.3. *Consider any upper semicontinuous, concave f with $R(f) > 0$ and $p^* = \sup f \in \mathbb{R}_{++}$ attained at $x^* \in \mathcal{E}$. Fixing $\alpha_k = (f^\Gamma(y_k) - d^*)/\|\zeta_k'\|^2$, if f satisfies the Łojasiewicz condition* (29) *with exponent $\theta > 0$, then the radial subgradient method (Algorithm 1) has $x_k = y_k/f^\Gamma(y_k)$ satisfy*

$$T \geq O\left(1/\epsilon^{2\theta}\right) \quad\implies\quad \min_{k<T}\left\{\frac{p^* - f(x_k)}{f(x_k)}\right\} \leq \epsilon.$$

Proof. By Proposition 3.4, the Łojasiewicz condition (28) holds at the dual minimizer $y^* = x^*/p^*$ for some constants r', C' with the same exponent θ. Integrating this condition (as done in [6, Theorem 5]) ensures every $y \in B(y^*, r')$ has the following local error bound

$$f^\Gamma(y) - d^* \geq \left(C'(1-\theta)\|y - y^*\|\right)^{1/(1-\theta)}. \tag{34}$$

The subgradient method must have some y_{k_0} in the ball $B(y^*, r')$ with

$$k_0 \leq \left(\frac{\|y_0 - y^*\|}{(C'(1-\theta)r')^{1/(1-\theta)}R(f)}\right)^2$$

since (33) ensures the average iterate has objective gap squared at most $(C'(1-\theta)r')^{2/(1-\theta)}$. Notice that the Polyak stepsize ensures the distance from the iterates y_k to y^* is nonincreasing as

$$\begin{aligned}\|y_{k+1} - y^*\|^2 &= \|y_k - y^*\|^2 - 2\alpha_k\zeta_k'^T(y_k - y^*) + \alpha_k^2\|\zeta_k'\|^2 \\ &\leq \|y_k - y^*\|^2 - 2\alpha_k(f^\Gamma(y_k) - d^*) + \alpha_k^2\|\zeta_k'\|^2 \\ &\leq \|y_k - y^*\|^2 - \frac{(f^\Gamma(y_k) - d^*)^2}{\|\zeta_k'\|^2} \leq \|y_k - y^*\|^2.\end{aligned}$$

Hence all $k \geq k_0$ have $y_k \in B(y^*, r')$ as well. Then our claimed convergence rate follows by bounding the number of iterations required to ensure the objective gap halves $f^\Gamma(y_{k_0+k}) - d^* \leq (f^\Gamma(y_{k_0}) - d^*)/2$. Applying the local error bound (34) to (33) initialized at y_{k_0} implies

$$\frac{1}{T}\sum_{k=0}^{T-1}(f^\Gamma(y_{k_0+k}) - d^*)^2 \leq \frac{(C'(1-\theta))^2(f^\Gamma(y_{k_0}) - d^*)^{2(1-\theta)}}{C^2R(f)^2T}.$$

Therefore some $k \leq 4\left(\frac{C'(1-\theta)}{R(f)}\right)^2/(f^\Gamma(y_{k_0}) - d^*)^{2\theta}$ iterations after k_0, the radially dual objective gap must have halved. Repeatedly applying this gives the following geometric sum limiting the number of iterations required to reach any $\epsilon' > 0$ level of radial dual accuracy

$$T \geq k_0 + \sum_{i=1}^{\infty} 4\left(\frac{C'(1-\theta)}{R(f)}\right)^2 \frac{1}{(2^i\epsilon')^{2\theta}} \implies \min_{k<T}\left\{f^\Gamma(y_k) - d^*\right\} \leq \epsilon'.$$

Selecting $\epsilon' = \epsilon/p^*$ gives the claimed result as

$$T \geq k_0 + \frac{4}{1-2^{2\theta}}\left(\frac{C'(1-\theta)}{R(f)}\right)^2\left(\frac{2p^*}{\epsilon}\right)^{2\theta} \implies \min_{k<T}\left\{\frac{p^*-f(x_k)}{f(x_k)}\right\} \leq \epsilon. \qquad \square$$

The previous pair of convergence theorems relied on using a Polyak stepsize, which requires the often impractical knowledge of d^*. This can be remedied by replacing the simple subgradient method in Algorithm 1 with a more sophisticated stepping scheme like [22] or restarting scheme like [46, 41, 40] which all attain similar convergence guarantees.

4.2 Radial Smoothing Method

Now we turn our attention to the radial smoothing method previously defined as Algorithm 2 in the context of smoothing the radial dual of our quadratic program. More generally, we consider primal problems maximizing a minimum of smooth functions over polyhedral constraints

$$p^* = \begin{cases} \max_x & \min\{f_j(x) \mid j = 1,\dots,m_1\} \\ \text{s.t.} & a_i^T x \leq b_i \quad \text{for } i = 1,\dots,m_2 \end{cases} \tag{35}$$

where each f_j is twice continuously differentiable and concave with $R(f_j) \geq R > 0$ and $D(f_j) \leq D < \infty$ and each $b_i > 0$. Note that having $R(f_j) > 0$ and $b_i > 0$ can be attained without loss of generality by translating a strictly feasible point in the domain of each f_j to the origin. Further, assuming $D(f_j) < \infty$ implies each f_j has bounded level sets and so each f_j is L-smooth on the level set $\{x \mid f_j(x) > 0\}$ for some $\sup\{\|\nabla^2 f_j(x)\| \mid f(x) > 0\} \leq L < \infty$. This objective is strictly upper radial and its radial dual is

$$d^* = \min_{y\in\mathcal{E}} \max\big\{f_j^\Gamma(x), (a_i/b_i)^T y \mid j \in \{1,\dots,m_1\}, i \in \{1,\dots,m_2\}\big\}. \tag{36}$$

Then we consider the smoothing of this objective for any $\eta > 0$ given by

$$g_\eta(y) = \eta \log\left(\sum_{j=1}^{m_1} \exp\left(\frac{f_j^\Gamma(y)}{\eta}\right) + \sum_{i=1}^{m_2} \exp\left(\frac{a_i^T y}{b_i \eta}\right)\right).$$

Our radial smoothing method (Algorithm 2) proceeds by minimizing this smoothing with Nesterov's accelerated method to produce a radially dual solution with accuracy $O(\eta)$. Nearly any other fast iterative method could be employed here instead, which could then avoid needing knowledge of problem constants. Converting this radial dual guarantee back to the primal problem gives the following primal convergence theorem.

Theorem 4.4. *Consider any problem of the form* (35). *Fixing* $L_\eta = (1+D/R)^3 L + \frac{\max\{1/R^2, \|a_i/b_i\|\}}{\eta}$ *and* $x_0 = 0$, *the radial smoothing method (Algorithm 2) has* $x_k = y_k/\max\{f_j^\Gamma(x), (a_i/b_i)^T y\}$ *feasible with*

$$\frac{p^* - \min\{f_j(x_k)\}}{\min\{f_j(x_k)\}} \leq \frac{2L_\eta(1+\eta p^* \log(m_1+m_2))^2 D^2}{p^*(k+1)^2} + \eta p^* \log(m_1+m_2).$$

In particular, setting $\eta = \epsilon/2\log(m_1+m_2)$, *this ensures the following* $O(1/\epsilon)$ *convergence rate*

$$k+1 \geq 2(1+p^*\epsilon/2)D\sqrt{\frac{(1+D/R)^3 L}{p^*\epsilon} + \frac{2\max\{1/R^2, \|a_i/b_i\|^2\}\log(m_1+m_2)}{p^*\epsilon^2}}$$

$$\implies \quad \frac{p^* - \min\{f_j(x_k)\}}{\min\{f_j(x_k)\}} \leq p^*\epsilon.$$

Proof. Observe that all of the m_1+m_2 functions defining g_η are convex (by (15)), $\max\{1/R, \|a_i/b_i\|\}$-Lipschitz continuous (by Proposition 3.1) and $(1+D/R)^3L$-smooth (by Corollary 3.3). Then [3, Proposition 4.1] ensures g_η is convex, is $(1+D/R)^3L+\frac{\max\{1/R^2,\|a_i/b_i\|\}}{\eta}$-smooth, and closely follows the radially dual objective with every $y\in\mathcal{E}$ satisfying

$$0 \le g_\eta(y) - \max\{f_j^\Gamma(y), (a_i/b_i)^Ty\} \le \eta\log(m_1+m_2). \tag{37}$$

Note that for any $s>0$, the corresponding primal objective super-level set is bounded by

$$\sup\{\|x\| \mid f_j(x)\ge s,\ a_i^Tx\le b_i\} \le D.$$

Then the bijection epi $f^\Gamma = \Gamma(\text{hypo } f)$ from (18) bounds every sub-level set of the dual with

$$\sup\{\|y\| \mid f_j^\Gamma(y)\le 1/s,\ (a_i/b_i)^Ty\le 1/s\} \le D/s.$$

In particular considering $s=p^*=1/d^*$ shows every radial dual minimizer has norm bounded by d^*D. Then the upper bound from (37) ensures the $d^*+\eta\log(m_1+m_2)$ sub-level set of g_η is nonempty and the lower bound from (37) allows us to bound this level set by

$$\sup\{\|y\| \mid g_\eta(y)\le d^*+\eta\log(m_1+m_2)\} \le (d^*+\eta\log(m_1+m_2))D$$

Therefore the distance from $y_0=0$ to a minimizer of g_η is at most $(d^*+\eta\log(m_1+m_2))D$.

Since g_η is smooth and has a minimizer, applying the standard accelerated method convergence guarantee [32] guarantees the iterates of our radial smoothing method have

$$g_\eta(y_k) - \inf g_\eta \le \frac{2L_\eta(d^*+\eta\log(m_1+m_2))^2D^2}{(k+1)^2}.$$

Converting this guarantee to be in terms of our radially dual objective, (37) ensures

$$\max\{f_j^\Gamma(y_k), (a_i/b_i)^Ty_k\} - d^* \le \frac{2L_\eta(d^*+\eta\log(m_1+m_2))^2D^2}{(k+1)^2} + \eta\log(m_1+m_2).$$

Finally, stating this to be in terms of the primal solution $x_k = y_k/\max\{f_j^\Gamma(x), (a_i/b_i)^Ty\}$ yields

$$\frac{p^*-\min\{f_j(x_k)\}}{\min\{f_j(x_k)\}} \le \frac{2L_\eta(1+\eta p^*\log(m_1+m_2))^2D^2}{p^*(k+1)^2} + \eta p^*\log(m_1+m_2). \qquad \square$$

Renegar [39] uses the same general technique to give accelerated convergence guarantees for solving the broad family of hyperbolic programming problems (which includes semidefinite programming) where the radial dual also admits a natural smoothing. The restarting schemes of [41] and [40] both explicitly consider restarting smoothing methods to attain improved convergence when growth conditions like the Łojasiewicz condition (28) hold. Due to Proposition 3.4, applying these more sophisticated methods to solve the radially dual problem will give rise to radial algorithms that enjoy the same improved convergence. The analysis of such a method should follow similarly to that of Theorem 4.3.

4.3 Radial Accelerated Method

Motivated by our example transforming the Poisson likelihood problem (9), algorithms can be designed to take advantage of the radial transformation extending smoothness on a level set to hold globally. Consider maximizing any twice differentiable concave function $f : \mathcal{E} \to \mathbb{R} \cup \{-\infty\}$ with bounded level sets. Then, without loss of generality, we have $0 \in \text{int}\ \{x \mid f(x) > 0\}$ and so f_+ is strictly upper radial. Letting $L = \sup\{\|\nabla^2 f(x)\| \mid f(x) > 0\}$, Corollary 3.3 ensures f_+^Γ is $(1 + D(f)/R(f))^3 L$-smooth on all of $\mathcal{E}$. Hence f_+^Γ can be minimized directly using Nesterov's accelerated method, giving the following *radial accelerated method* defined by Algorithm 3.

Algorithm 3 The Radial Accelerated Method

Require: $f : \mathcal{E} \to \overline{\mathbb{R}}_{++}$, $x_0 \in \text{dom}\ f$, $L > 0$, $T \geq 0$

$(y_0, v_0) = \Gamma(x_0, f(x_0))$ and $\tilde{y}_0 = y_0$ — *Transform into the radial dual*
for $k = 0 \ldots T-1$ **do**
  $\tilde{y}_{k+1} = y_k - \nabla f^\Gamma(y)/(1 + D(f)/R(f))^3 L$ — *Run the accelerated method*
  $y_{k+1} = \tilde{y}_{k+1} + \frac{k-1}{k+2}(\tilde{y}_{k+1} - \tilde{y}_k)$
end for
$(x_T, u_T) = \Gamma(y_T, f^\Gamma(y_T))$ — *Transform back to the primal*
return x_T

This radial algorithm inherits the primal accelerated method's $O(\sqrt{L\text{dist}(x_0, X^*)^2/\epsilon})$ rate, only requiring L-smoothness on the level set $\{x \mid f(x) > 0\}$ as follows.

Theorem 4.5. *Consider any twice differentiable, concave f with $R(f) > 0$, $D(f) < \infty$, $L = \sup\{\|\nabla^2 f(x)\| \mid f(x) > 0\}$, and $p^* = \sup f \in \mathbb{R}_{++}$ attained on some set $X^* \subseteq \mathcal{E}$. Fixing $x_0 = 0$, the radial accelerated method (Algorithm 3) has for any $\epsilon > 0$,*

$$k + 1 \geq (1 + D(f)/R(f))^{3/2} \sqrt{\frac{2L\text{dist}(x_0, X^*)^2}{p^*\epsilon}} \implies \frac{p^* - f(x_k)}{f(x_k)} \leq \epsilon.$$

Proof. Recall the f^Γ is convex by (15) and is $(1 + D(f)/R(f))^3 L$-smooth by Corollary 3.3. Then Nesterov's classic analysis [32] ensures our radially dual iterates converge with

$$f^\Gamma(y_k) - d^* \leq \frac{2(1 + D(f)/R(f))^3 L\text{dist}(y_0, Y^*)^2}{(k+1)^2}$$

where $Y^* = X^*/p^*$. Letting $(x_k, u_k) = \Gamma(y_k, v_k)$ yields primal iterates with $f(x_k) \geq 1/f^\Gamma(y_k)$. Then multiplying this bound through by $1/d^* = p^*$ produces the primal guarantee

$$\frac{p^* - f(x_k)}{f(x_k)} \leq \frac{2(1 + D(f)/R(f))^3 L\text{dist}(y_0/d^*, X^*)^2}{p^*(k+1)^2}.$$

Noting that $y_0/d^* = x_0 = 0$, this gives the claimed convergence guarantee. □

A few remarks on this convergence result. The additional coefficient of $(1 + D(f)/R(f))^{3/2}$ is quite pessimistic as many of the examples we have considered have radial dual smoother than the primal, but Corollary 3.3 fails to capture this potential upside in its $O(L)$ bound. For particular applications, we expect much tighter bounds on the radially dual smoothness are possible. The proposed radial accelerated method unrealistically relies on knowledge of our smoothness constant upper bound $(1 + D(f)/R(f))^3 L$. However, this can be remedied by including a line-search/backtracking as done in [2, 34].

Under growth conditions, the convergence of accelerated methods also improves. For example, applying the adaptive accelerated gradient method of [27] to solve the radially dual problem would give a radial method that speeds up in the presence of primal growth conditions by Proposition 3.4. The analysis of such a method should follow similarly to that of Theorem 4.3.

5 Radial Algorithms for Nonconcave Maximization

Our radial duality theory applies beyond the concave maximization problems that have been considered so far. The foundational theorem (1) establishes that our radial duality applies to the broader family of upper radial functions.

5.1 Examples of Radial Duality with Nonconvex Objectives or Constraints

Geometrically, upper radial functions all have a star-convex hypograph with respect to the origin [17, Lemma 3.1], meaning that all $(y, v) \in \text{hypo } f$ have $(y, v)/t \in \text{hypo } f$ for all $t \geq 1$. Star-convexity has been considered throughout the optimization literature. The structure of optimizing over star-convex constraint sets has been considered as early as [43]. Efficient global optimization of star-convex objectives is possible if star-convexity holds with respect to a global optimizer (see [35, 18, 26, 19, 21]). However, in general, even linear optimization over star-convex bodies is NP-hard [8]. Our model of star-convexity w.r.t. the origin captures this NP-hard case.

5.1.1. Star-Convex Constraints We say that a set $S \subseteq \mathcal{E}$ is *star-convex with respect to the origin* if every $x \in S$ has the line segment $\lambda x \in S$ for all $0 \leq \lambda \leq 1$. This is exactly the condition needed to ensure the indicator function $\iota_S(x) = \begin{cases} +\infty & \text{if } x \in S \\ 0 & \text{if } x \notin S. \end{cases}$ is strictly upper radial[6]. Then the radial dual of such a star-convex set's indicator function is given by the gauge

$$\iota_S^\Gamma(y) = \sup\{v > 0 \mid v \cdot \iota_S(y/v) \leq 1\} = \sup\{v > 0 \mid y/v \notin S\} = \inf\{\lambda > 0 \mid y \in \lambda S\} = \gamma_S(y).$$

Importantly, the gauge $\gamma_S(y)$ is convex if and only if S is convex. As a result, algorithms utilizing the radial dual of star-convex constraints avoid needing difficult nonconvex orthogonal projections, replacing them with evaluating a nonconvex gauge function appearing in the objective.

One important example where star-convex sets arises comes from considering chance constraints [23, 31, 48]. Given some distribution over potential constraint sets $S_\xi \subseteq \mathcal{E}$, a robust problem formulation may want to ensure that the constraint is satisfied with probability at least $\Lambda \in [0, 1]$. Then the chance-constrained feasible region is $S = \{x \mid \mathbb{P}(x \in S_\xi) \geq \Lambda\}$. If each potential constraint set is convex with $0 \in S_\xi$, then the chance-constrained set S is star-convex w.r.t. the origin.

5.1.2. Optimization over Compact Sets Now we generalize our previous example from Section 2 where we saw that any nonconcave quadratic program with a compact polyhedral feasible region could be rescaled for our radial duality to apply. Consider maximizing any continuously

[6]This is essentially by definition as $v \cdot \iota_S(y/v)$ is nondecreasing in v if and only if S is star-convex w.r.t. the origin. Then its simple to check this function is upper semicontinuous and is vacuously strictly increasing on its effective domain $\text{dom } \iota_S = \emptyset$, which is empty.

differentiable function f over a compact set S that is star-convex w.r.t. the origin. Supposing $f(0) > 0$, this is equivalent to the following maximization problem of the primal form (2)

$$\max_{y\in\mathcal{E}} \min\{(1+\lambda f(x))_+, \iota_S(x)\}$$

for any $\lambda > 0$. We can check when this objective is strictly upper radial (and so our duality holds) by considering whether its perspective function is strictly increasing on its domain:

$$v \cdot \min_i\{(1+\lambda f(y/v))_+,\ \iota_S(y/v)\} = \begin{cases} (v+\lambda v f(y/v))_+ & \text{if } y/v \in S \\ 0 & \text{otherwise.} \end{cases}$$

The partial derivative of this with respect to v at any $y/v \in S \cap \mathrm{dom}\ (1+\lambda f)_+$ is

$$1 - \lambda(\nabla f(y/v), -1)^T(y/v, f(y/v)).$$

Noting that $(\nabla f(x), -1)^T(x, f(x))$ is continuous on the compact set $S \cap \mathrm{dom}\ (1+\lambda f)_+$, we can select $\lambda > 0$ small enough to always have $1 - \lambda(\nabla f(y/v), -1)^T(y/v, f(y/v)) > 0$. Doing so makes our objective strictly upper radial and hence our radial duality applies.

5.1.3. Nonconvex Regularization Many optimization tasks take the additive composite form

$$\max_{y\in\mathcal{E}} f(x) - r(x)$$

where f is an upper semicontinuous, concave function with $f(0) > 0$ and $r(x)$ is an added (or rather subtracted since we are maximizing) regularization term. Many sparsity inducing regularization penalties decompose as a sum over the x's coordinates $r(x) = \sum_{i=1}^n \sigma(x_i)$ for some simple nonconvex function $\sigma : \mathbb{R} \to \mathbb{R}$. For example, ℓ_q-regularization sets $\sigma(t) = \lambda|t|^q$ for some $0 < q < 1$, bridging the gap between ℓ_0 and ℓ_1-regularization, and smoothly clipped absolute deviation (SCAD) regularization [12] sets

$$\sigma(t) = \begin{cases} \lambda|t| & \text{if } |t| \le \lambda \\ (-|t|^2 + 2a\lambda|t| - \lambda^2)/2(a-1) & \text{if } \lambda < |t| \le a\lambda \\ (1+a)\lambda^2/2 & \text{if } |t| > a\lambda. \end{cases}$$

for some constants $a > 2$ and $\lambda > 0$. Many more regularizers are of this form, like MCP [49] and firm thresholding [15]. See [45] for a survey of numerous other important nonconvex regularization formulations and their usage in practice.

These regularizers are all continuous and have $r(y/v)$ nonincreasing in v. These two simple properties suffice to guarantee subtracting r from f will not break its upper radiality since

$$v(f(y/v) - r(y/v))_+ = \max\{vf(y/v) - vr(y/v), 0\}$$

is a sum of two upper semicontinuous, nondecreasing functions in v. As a result, our radial duality applies to the potentially nonconcave primal objective $(f(x) - r(x))_+$.

5.1.4. Optimization with Outliers Many learning problems take the form of minimizing a stochastic loss function $\mathbb{E}_\xi[f(x,\xi)]$ using a finite sample approximation. Given i.i.d. samples $\xi_1,\dots,\xi_s$, this problem can be formulated as the following maximization

$$\max_{x\in\mathcal{E}} \frac{1}{s}\sum_{i=1}^{s} -f(x,\xi_i).$$

If each $-f(\cdot,\xi_i)$ is concave, a translation will ensure every $-f(\cdot,\xi_i)$ is upper radial and our radial duality can be applied. In the presence of t outliers in the s samples $\xi_1,\dots,\xi_s$, this finite sample approximation could be improved to only consider the loss function on the best $s-t$ samples

$$\max_{x\in\mathcal{E}} \max\left\{\frac{1}{s-t}\sum_{i\in S} -f(x,\xi_i) \mid S\subseteq\{1...s\}, |S| = s-t\right\}.$$

Provided each $-f(\cdot,\xi_i)$ is upper radial, this whole objective will be upper radial by [17, Corollary 3.11] and so our radial duality applies. The minimax formulation of [47] exactly corresponds to this problem formulation at its equilibrium. By the same corollary, our radial duality also applies to maximizing the $(s-t)$th largest element of $\{-f(x,\xi_i)\}_{i=1}^s$. Such an optimization problem captures the classic idea of least median of squares regression [42].

5.2 Example Nonconcave Guarantee for the Radial Subgradient Method

In this concluding section, we demonstrate the style of results possible from applying our radial duality to nonconcave maximization. In particular, we consider the nonconcave, nonsmooth primal problem of maximizing the minimum of a set of twice continuously differentiable, strictly upper radial f_j over some convex set $S\subseteq\mathcal{E}$

$$p^* = \begin{cases}\max_x & \min\{f_j(x)\mid j=1,\dots,m\}\\ \text{s.t.} & x\in S\end{cases} = \max_{x\in\mathcal{E}}\min\{f_j(x),\iota_S(x)\} \tag{38}$$

where each f_j has $R(f_j)\geq R>0$ and bounded level sets $D(f_j)\leq D<\infty$ and the origin lies in the interior of the constraint set $B(0,R)\subseteq S$. Let $L\geq \sup\{\|\nabla^2 f_j(x)\| \mid f_j(x)>0, x\in S\}$ bound the smoothness of each f_j on this compact level set.

This primal is strictly upper radial since each function defining the minimum is strictly upper radial and so our radial duality applies. The radial dual of this problem is

$$d^* = \min_{y\in\mathcal{E}}\max\{f_j^\Gamma(y), \gamma_S(y)\}. \tag{39}$$

Note each $f_j^\Gamma(y)$ is convex if and only if f_j is concave by (15). Hence if our primal (38) is nonconcave, our radial dual (39) will be nonconvex. Regardless, our previously proposed radial subgradient method (Algorithm 1) can still be applied and analyzed.

Recently, convergence theory for subgradient methods without convexity has been developed. Particularly, consider minimizing a nonconvex, nonsmooth function $g\colon\mathcal{E}\to\mathbb{R}$ that is bounded below. Then [10, Theorem 3.1] ensures that provided g is uniformly M-Lipschitz and ρ-weakly convex (defined as $g+\frac{\rho}{2}\|\cdot\|^2$ being convex), the subgradient method $y_{k+1}=y_k-\epsilon\zeta_k/\|\zeta_k\|^2$ for $\zeta_k\in\partial_P g(y_k)$ will have some y_k be nearly stationary on the Moreau envelope of g. In particular, this implies some y_k will have a nearby y that is nearly stationary

$$T\geq\frac{\rho M^2(g(y_0)-\inf g)}{\epsilon^4} \implies \min_{k<T}\{\|y-y_k\|\}\leq\frac{\epsilon}{2\sqrt{\rho}} \text{ with } \operatorname{dist}(0,\partial_P g(y))\leq\sqrt{\rho}\epsilon. \tag{40}$$

Applying this machinery on the radial dual allows us to ensure a nearly stationary point y near a dual iterate y_k exists. Then converting this guarantee back to the primal gives the following primal convergence guarantee, preserving the above $O(1/\epsilon^4)$ rate despite not assuming the primal (38) is either Lipschitz or weakly convex.

Theorem 5.1. *Consider any problem of the form* (38) *with* $p^* \in \mathbb{R}_{++}$. *Fixing* $x_0 = 0$ *and* $\alpha_k = \epsilon/\|\zeta_k'\|^2$, *the radial subgradient method (Algorithm 1) has* $x_k = y_k/\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}$ *satisfy*

$$\begin{aligned} T \geq \frac{(1+D/R)^3 L(\min\{f_j(x_0)\} - p^*)}{R^2 \min\{f_j(x_0)\} p^* \epsilon^4} \\ \implies \min_{k<T}\{\|x - x_k\|\} \leq \frac{p^*\epsilon}{2\sqrt{(1+D/R)L}} \\ \textit{with } \operatorname{dist}(0, \partial^P \min\{f_j, \iota_{a_i^T x \leq b_i}\}(x)) \leq \frac{p^*\sqrt{(1+D/R)^3 L}\epsilon}{1 - \sqrt{(1+D/R)^3 L}\epsilon D} \end{aligned}$$

for some nearby $x \in \mathcal{E}$ *provided* $0 < \epsilon < 1/\sqrt{(1+D/R)^3 L}D$.

Proof. Observe that each function in the maximum defining the radial dual (36) is $1/R$-Lipschitz (by Proposition 3.1) and each f_j^Γ is $(1+D/R)^3 L$-smooth (by Corollary 3.3). Then the whole radially dual objective $\max\{f_j^\Gamma(y), \gamma_S(y)\}$ is $1/R$-Lipschitz and $(1+D/R)^3 L$-weakly convex. Hence even though our primal is not assumed to be either Lipschitz or weakly convex, these two properties occur in the radial dual due to each f_i having $R(f_i) > 0$ and smoothness on the level set $\{x \mid f_j(x) > 0\}$ respectively. Then we can apply (40) implying a nearby dual solution y has

$$T \geq \frac{(1+D/R)^3 L(\min\{f_j^\Gamma(y_0)\} - d^*)}{R^2\epsilon^4} \implies \min_{k<T}\{\|y - y_k\|\} \leq \frac{\epsilon}{2\sqrt{(1+D/R)^3 L}}$$
$$\text{with } \operatorname{dist}(0, \partial_P \max\{f_j^\Gamma, \gamma_S\}(y)) \leq \sqrt{(1+D/R)^3 L}\ \epsilon.$$

Relating this guarantee to the primal is done in the following two steps. First, we show the nearby radial dual solution y corresponds to a primal solution $x = y/\max\{f_j^\Gamma(y), \gamma_S(y)\}$ that is also near the primal iterates $x_k = y_k/\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}$. Then relating the dual stationarity of y to the primal stationarity of x completes our proof, showing it is a nearby, nearly stationary primal solution.

Observe that having $\|y - y_k\| \leq \epsilon/2\sqrt{(1+D/R)^3 L}$ ensures the distance $\|x - x_k\|$ is at most

$$\begin{aligned} \|x - x_k\| &= \left\| \frac{y}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} - \frac{y_k}{\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}} \right\| \\ &\leq \frac{\|y - y_k\|}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} + \left\| \frac{y_k}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} - \frac{y_k}{\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}} \right\| \\ &= \frac{\|y - y_k\|}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} + \|x_k\| \left| \frac{\max\{f_j^\Gamma(y_k), \gamma_S(y_k)\}}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} - 1 \right| \\ &\leq \frac{\|y - y_k\|}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} + \frac{D\|y - y_k\|/R}{\max\{f_j^\Gamma(y), \gamma_S(y)\}} \\ &\leq \frac{1+D/R}{d^*}\|y - y_k\| \leq \frac{\epsilon}{2d^*\sqrt{(1+D/R)L}} \end{aligned}$$

where the first inequality uses the triangle inequality, the second uses the bounded primal level sets and the radially dual $1/R$-Lipschitz continuity, and the third uses that $d^* = 1/p^* \in \mathbb{R}_{++}$.

Lastly, let $v = \max\{f_j^\Gamma(y), \gamma_S(y)\}$, $u = 1/v$ and $\zeta' \in \partial_P \max\{f_j^\Gamma, \gamma_S\}(y)$ denote a radially dual subgradient with $\|\zeta'\| \leq \sqrt{(1+D/R)^3 L}\ \epsilon$. Then we can bound

$$(\zeta', -1)^T(y, v) \leq \|\zeta'\|\|y\| - v \leq \sqrt{(1+D/R)^3 L\epsilon}\|x\|/u - 1/u \leq -(1 - \sqrt{(1+D/R)^3 L\epsilon D})/p^* < 0.$$

Noting that $(\zeta', -1) \in N^P_{\text{epi } \min\{f_j^\Gamma, \gamma_S\}}(y, v)$, the primal then has a supgradient

$$\zeta := \frac{\zeta'}{(\zeta', -1)^T(y, v)} \in \partial^P \min\{f_j, \iota_S\}(x)$$

by (24) applied to the radial dual. This primal subgradient has norm at most $O(\epsilon)$ since

$$\|\zeta\| = \left\|\frac{\zeta'}{(\zeta', -1)^T(y, v)}\right\| = \frac{\|\zeta'\|}{|(\zeta', -1)^T(y, v)|} \leq \frac{p^*\sqrt{(1+D/R)^3 L\epsilon}}{1 - \sqrt{(1+D/R)^3 L\epsilon D}}. \qquad \square$$

Acknowledgments. The author thanks Jim Renegar broadly for inspiring this work and concretely for providing feedback an early draft and Rob Freund for constructive thoughts helping focus this work.

References

[1] Heinz H. Bauschke, Jerome Bolte, and Marc Teboulle. A descent lemma beyond lipschitz gradient continuity: First-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2017.

[2] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[3] Amir Beck and Marc Teboulle. Smoothing and first order methods: A unified framework. *SIAM J. Optim.*, 22:557–580, 2012.

[4] Mario Bertero, Patrizia Boccacci, Gabriele Desiderà, and Giuseppe Vicidomini. Image deblurring with poisson data: from cells to galaxies. *Inverse Problems*, 25(12):123006, nov 2009.

[5] Jérôme Bolte, Aris Daniilidis, and Adrian Lewis. The łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM Journal on Optimization*, 17(4):1205–1223, 2007.

[6] Jérôme Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce W. Suter. From error bounds to the complexity of first-order descent methods for convex functions. *Mathematical Programming*, 165(2):471–507, Oct 2017.

[7] James V. Burke and Michael C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[8] Karthekeyan Chandrasekaran, Daniel Dadush, and Santosh Vempala. *Thin Partitions: Isoperimetric Inequalities and a Sampling Algorithm for Star Shaped Bodies*, pages 1630–1645.

[9] Francis H. Clarke, Yuri S. Ledyaev, Ronald J. Stern, and Peter R. Wolenski. *Nonsmooth Analysis and Control Theory.* Springer-Verlag, Berlin, Heidelberg, 1998.

[10] Damek Davis and Dmitriy Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

[11] William S. Dorn. Duality in quadratic programming. *Quarterly of Applied Mathematics*, 18(2):155–162, 1960.

[12] Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

[13] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(12):95–110, 1956.

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

[15] Hong-Ye Gao and Andrew G. Bruce. Wave shrink with firm shrinkage. *Statistica Sinica*, 7(4):855 – 874, 1997.

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

[17] Benjamin Grimmer. Radial Duality Part I: Foundations. *arXiv:2104.11179*, April 2021.

[18] Sergey Guminov and Alexander Gasnikov. Accelerated Methods for α-Weakly-Quasi-Convex Problems. *arXiv e-prints*, October 2017.

[19] Sergey Guminov, Yurii Nesterov, Pavel Dvurechensky, and Alexander Gasnikov. Accelerated primal-dual gradient descent with linesearch for convex, nonconvex, and nonsmooth optimization problems. *Dokl. Math.*, 99:125–128, 2019.

[20] Niao He, Zaïd Harchaoui, Yichen Wang, and Le Song. Fast and simple optimization for poisson likelihood models. *CoRR*, abs/1608.01264, 2016.

[21] Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1894–1938. PMLR, 09–12 Jul 2020.

[22] Patrick R. Johnstone and Pierre Moulin. Faster subgradient methods for functions with hölderian growth. *Math. Program.*, 180(1):417–450, 2020.

[23] Willem K. Klein Haneveld, Maarten H. van der Vlerk, and Ward Romeijnders. *Chance Constraints*, pages 115–138. Springer International Publishing, Cham, 2020.

[24] Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. *Annales de l'institut Fourier*, 48(3):769–783, 1998.

[25] Simon Lacoste-Julien, Mark Schmidt, and Francis R. Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method. *CoRR*, abs/1212.2002, 2012.

[26] Jasper C.H. Lee and Paul Valiant. Optimizing star-convex functions. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 603–614, 2016.

[27] Mingrui Liu and Tianbao Yang. Adaptive accelerated gradient converging method under holderian error bound condition. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[28] Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

[29] Stanislaw Lojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.

[30] Mahesh Chandra Mukkamala, Jalal Fadili, and Peter Ochs. Global convergence of model function based bregman proximal minimization algorithms, arXiv:2012.13161, 2020.

[31] Arkadi Nemirovski and Alexander Shapiro. Convex approximations of chance constrained programs. *SIAM Journal on Optimization*, 17(4):969–996, 2007.

[32] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[33] Yurii Nesterov. Smooth minimization of non-smooth functions. *Math. Program.*, 103(1):127152, May 2005.

[34] Yurii Nesterov. Universal gradient methods for convex optimization problems. *Math. Program.*, 152(1-2):381–404, August 2015.

[35] Yurii Nesterov and Boris Polyak. Cubic regularization of newton method and its global performance. *Math. Program.*, 108:177–205, 08 2006.

[36] Boris T. Polyak. Minimization of unsmooth functionals. *USSR Computational Mathematics and Mathematical Physics*, 9(3):14–29, 1969.

[37] Boris T. Polyak. Sharp minima. *Institute of Control Sciences Lecture Notes,Moscow, USSR. Presented at the IIASA Workshop on Generalized Lagrangians and Their Applications, IIASA, Laxenburg, Austria.*, 1979.

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

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

[40] James Renegar and Benjamin Grimmer. A Simple Nearly-Optimal Restart Scheme For Speeding-Up First Order Methods. *To appear in Foundations of Computational Mathematics*, 2021.

[41] Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart, and acceleration. *SIAM Journal on Optimization*, 30(1):262–289, 2020.

[42] Peter J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[43] A.M. Rubinov and A.A. Yagubov. The space of star-shaped sets and its applications in nonsmooth optimization. *Mathematical Programming Studies*, 29, 1986.

[44] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd. OSQP: an operator splitting solver for quadratic programs. *Mathematical Programming Computation*, 12(4):637–672, 2020.

[45] Fei Wen, Lei Chu, Peilin Liu, and Robert C. Qiu. A survey on nonconvex regularization-based sparse and low-rank recovery in signal processing, statistics, and machine learning. *IEEE Access*, 6:69883–69906, 2018.

[46] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *Journal of Machine Learning Research*, 19(6):1–33, 2018.

[47] Jin Yu, Anders Eriksson, Tat-Jun Chin, and David Suter. An adversarial optimization approach to efficient outlier removal. *J Math Imaging Vis*, 48:451466, 2014.

[48] Yuan Yuan, Zukui Li, and Biao Huang. Robust optimization approximation for joint chance constrained optimization problem. *J Glob Optim*, 67:805–827, 2017.

[49] Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894 – 942, 2010.