SIAM J. OPTIM.
Vol. 33, No. 1, pp. 36-62

SURVEY DESCENT: A MULTIPOINT GENERALIZATION OF GRADIENT DESCENT FOR NONSMOOTH OPTIMIZATION*

X. Y. HAN† AND ADRIAN S. LEWIS†

Abstract. For strongly convex objectives that are smooth, the classical theory of gradient descent ensures linear convergence relative to the number of gradient evaluations. An analogous nonsmooth theory is challenging. Even when the objective is smooth at every iterate, the corresponding local models are unstable, and the number of cutting planes invoked by traditional remedies is difficult to bound, leading to convergence guarantees that are sublinear relative to the cumulative number of gradient evaluations. We instead propose a multipoint generalization of the gradient descent iteration for local optimization. While our iteration was designed with general objectives in mind, we are motivated by a "max-of-smooth" model that captures the subdifferential dimension at optimality. We prove linear convergence when the objective is itself max-of-smooth, and experiments suggest a more general phenomenon.

Key words. first-order method, nonsmooth optimization, gradient descent, local linear convergence

MSC codes. 90C25, 65K05, 49M37

DOI. 10.1137/21M1468450

1. Introduction. To approach our target—a fast local iteration for nonsmooth optimization—let us recall, as motivation, the classical foundations of gradient descent (GD). For some function h, denote the linear approximation of h centered at some differentiable iterate $\tilde{x}$ by

$$\ell_{\tilde{x}}^{h}(x) \equiv h(\tilde{x}) + \nabla h(\tilde{x})^T (x - \tilde{x}),$$

where $\nabla h(\cdot)$ is the gradient of h. When h is smooth, given a step-size $\frac{1}{L}$, the canonical GD step $\tilde{x}^+ = \tilde{x} - \frac{1}{L}\nabla h(\tilde{x})$ produces the global minimum of the following local model:

$$h_{\tilde{x}}^{\mathrm{GD}}(x) = \ell_{\tilde{x}}^{h}(x) + \frac{L}{2}\|x - \tilde{x}\|_2^2. \tag{1.1}$$

The model $h_{\tilde{x}}^{\mathrm{GD}}$ incorporates the local linear behavior observed around $\tilde{x}$ through the gradient $\nabla h(\tilde{x})$ as well as the prior belief that the gradients themselves are L-Lipschitz functions of x. After iterating, one then builds a new model around $\tilde{x}^+$ and repeats the procedure. Indeed—on convex, L-smooth objectives—canonical results (as described in standard texts such as [2, Chapter 10]) guarantee that every GD iteration will both reduce the objective value and move closer to the true global minimizer of h.

When the Hessian is available, one can alternatively consider the Newton model

$$h_{\tilde{x}}^{\mathrm{Newton}}(x) = \ell_{\tilde{x}}^{h}(x) + \frac{1}{2}(x - \tilde{x})^T \nabla^2 h(\tilde{x})\,(x - \tilde{x}). \tag{1.2}$$

For first-order methods, however, the Hessian $\nabla^2 h(\tilde{x})$ is inaccessible. This motivates popular quasi-Newton approaches such as the Broyden–Fletcher–Goldfarb–Shanno

* Received by the editors December 30, 2021; accepted for publication (in revised form) September 27, 2022; published electronically January 19, 2023.
https://doi.org/10.1137/21M1468450
Funding: Research supported in part by National Science Foundation grant DMS-2006990.
† ORIE, Cornell University, Ithaca, NY 14850 USA (xh332@cornell.edu, adrian.lewis@cornell.edu).

(BFGS) algorithm (as described in standard texts such as [39, Chapter 6]) that minimizes an approximation to (1.2) after estimating $\nabla^2 h(\tilde{x})$ using only gradient differences and then performs a line search.

For nonsmooth objectives, both of the above smoothness-hypothesizing quadratic models—and their variants—are inadequate for generating theoretically guaranteed improvement at every step: They fail to capture the discontinuity of objective gradients. Thus, some algorithms work instead with a richer class of model functions. For example, one class of multipoint methods (discussed later in section 1.4) adaptively refines a lower cutting-plane model of the objective through a series of "null steps" around each iterate until it achieves descent in a "serious step." Such methods work well under reasonable conditions, finitely many null steps always sufficing to construct a successful serious step. However, analyzing convergence relative to all steps (null and serious) is challenging. Current bounds on the cumulative number of steps remain sublinear due to the difficulty of uniformly bounding the number of "in-between" null steps.

1.1. The survey descent iteration. In this work, we will propose and analyze a new local *survey descent* iteration for nonsmooth objectives h.

DEFINITION 1 (survey descent iteration $\{(P_i^{\mathcal{S}})\}_{i=1}^{k}$). *Given a survey of k points, $\mathcal{S} = \{s_i\}_{i=1}^{k}$, at which the nonsmooth objective h is differentiable and a step-control parameter L, for each $i = 1, \ldots, k$, define the ith subproblem $(P_i^{\mathcal{S}})$ as follows:*

$$\min_{x} \left\| x - \left(s_i - \frac{1}{L} \nabla h\left(s_i\right) \right) \right\|_2^2 \tag{1.3}$$

$$\text{s.t. } \ell_{s_j}^h(x) + \frac{L}{2} \left\| x - s_j \right\|_2^2 \leq \ell_{s_i}^h(x) \ \forall \ j \neq i. \tag{1.4}$$

We refer to the solving of all subproblems $\{(P_i^{\mathcal{S}})\}_{i=1}^{k}$ as a survey descent iteration.

When $(P_i^{\mathcal{S}})$ is feasible, we denote its optimal solution as s_i^+. When all subproblems are feasible, we say that the entire survey descent iteration is *feasible* and call $\mathcal{S}^+ = \{s_i^+\}_{i=1}^{k}$ the *outputs* of the iteration; otherwise, we say the entire iteration is *infeasible*. After a feasible iteration, we would update $\mathcal{S} \leftarrow \mathcal{S}^+$ and repeat the survey descent iteration on the updated survey.

Observe that, if $k = 1$, survey descent reduces to GD since there is then only one subproblem $(P_1^{\mathcal{S}})$ and survey point s_1, the constraints (1.4) are then empty, and the objective (1.3) of $(P_1^{\mathcal{S}})$ is then minimized by $s_1 - \frac{1}{L}\nabla h(s_1)$, which is the GD step from s_1. Thus, we can consider survey descent a generalization of GD.

1.1.1. Main results. When the function h is a maximum of k smooth functions, we prove local linear convergence of the survey descent iteration to the minimizer under reasonable conditions (Theorems 17–19). More precisely, we prove the following local properties of survey descent given input surveys sufficiently close to the minimizer of h:

- Survey descent iterations are always feasible.
- Survey descent outputs are always unique, and h remains differentiable at these outputs.
- Surveys converge Q-linearly to $\arg\min_x h(x)$ when survey descent is applied repeatedly.
- Survey function values converge R-linearly to $\min_x h(x)$ when survey descent is applied repeatedly.

Here, we follow the terminology of [41, Chapter 9]. A major assumption for our development and any practical implementation is the availability of a workable choice of the survey size, k, a question associated with the "active structure" of h at its minimizer. We discuss this choice throughout the exposition and present some empirical heuristics in Remark 20.

1.2. Linear convergence and nonsmooth objectives. In Figure 1, we illustrate a simple experiment on a max-of-smooth function objective suggesting linear convergence. Figure 2 suggests that this behavior persists in higher dimensions. Moreover, Figures 3 and 4, which apply survey descent on objectives not expressible as the

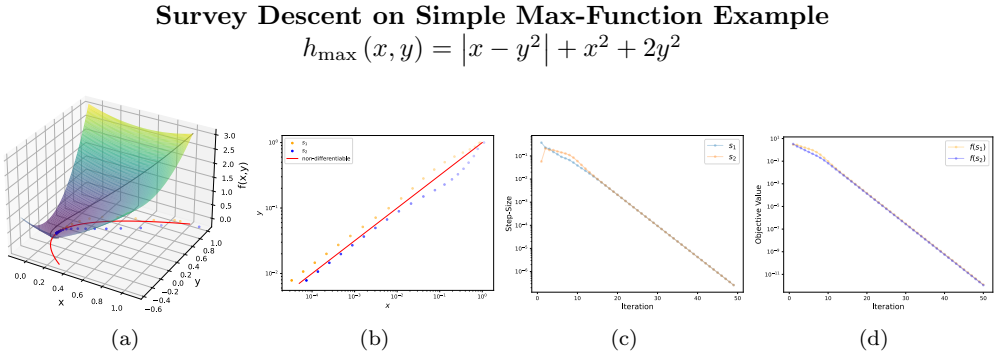


FIG. 1. *Survey descent iteration on max-function. We use survey descent iterations to minimize the nonsmooth objective* $h_{\max}(x,y)$ *that achieves minimum value* 0 *at unique minimizer* $\bar{x} = (0,0)$. *We use step-control parameter* $L = 10$ *and initialize with survey points* $s_1 = (0.9,1)$ *and* $s_2 = (1.1,1)$. *Panel* (a) *visualizes* $h_{\max}$. *The survey descent iterates (blue and orange dots) are shown in the xy-plane along with the* $x = y^2$ *curve (red) on which* $h_{\max}$ *is nondifferentiable. Panel* (b) *shows the location of iterates in the xy-plane where darker colors correspond to later iterations. Panel* (c) *shows the step-size (the magnitude of the difference between two consecutive iterates) at each iteration. Panel* (d) *shows the function value of the iterates. Iterates and function values converge linearly to global minimizer and minimum, respectively. Both survey points,* s_1 *and* s_2*, remain on the same smooth piece of the objective throughout the optimization. Sections 2–4 will theoretically derive these behaviors on objectives that are the maximum of smooth functions—of which* $h_{\max}$ *is a simple example.*

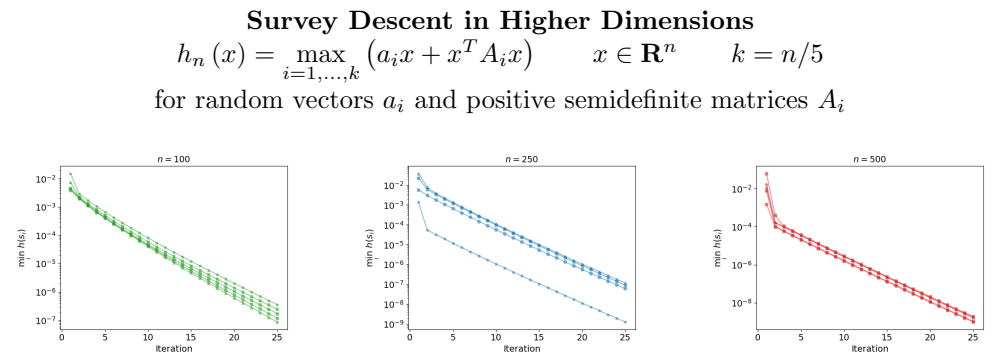


FIG. 2. *Panels plot the best objective values over* 25 *iterations, on* 5 *instances, for dimension* $n = 100, 250, 500$. *The initialization heuristic always chose the true number of components,* k*, as the survey size without a priori information. The data* $A_i = C_i^T C_i$ *and* a_i *were generated from square matrices* C_i *and vectors with standard Gaussian entries, the set* $\{a_i\}$ *being adjusted to ensure that its convex hull contains zero, giving the optimal solution* $x^* = 0$.

Survey Descent (with BFGS init.) on *Non*-Max-of-Smooth Objective

$$h_{\mathrm{ME}}(x,y,z,w) = \mathrm{MaxEigenvalue}\left(\begin{bmatrix} x & y & z \\ y & -x & w \\ z & w & -1 \end{bmatrix}\right)$$

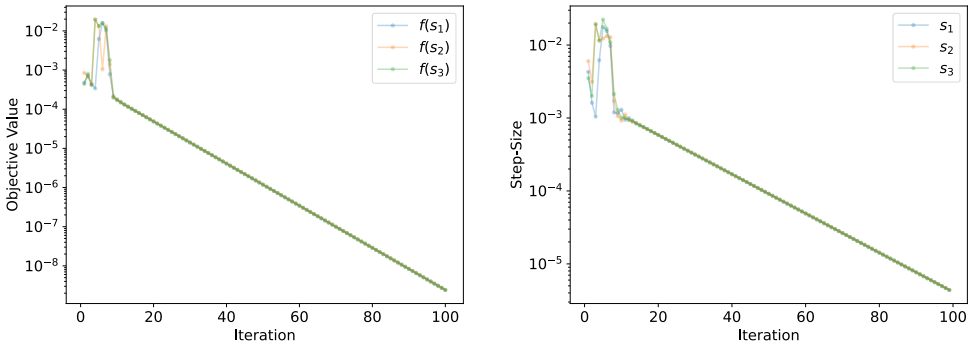


FIG. 3. *Survey descent of non-max-of-smooth objective. We use survey descent iterations with step-control parameter* $L = 10$ *to minimize* h_{ME} *with minimum value* 0 *at minimizer* $\bar{x} = (0,0,0,0)$. h_{ME} *is nonsmooth and not expressible as a maximum of smooth functions, as is clear by noting* $h_{ME}(x,y,0,0) = \sqrt{x^2+y^2}$. *To create an initializing survey, we run* 20 *preliminary iterations of BFGS exactly as described in* [31] *starting from* $(1,1,1,1)$. *We then chose both the size of a initializing survey—*3 *points—as well as the survey points themselves through an empirical heuristic based on the dimensionality of the BFGS iterate gradients. The left and right panels plot the objective values and step-sizes (the magnitude of the difference between two consecutive iterates), respectively, of each survey point during the survey descent iterations. After some initial fluctuations, we see the objective value and iterates of all survey descent survey-iterates converge at stably linear rates to the optimum. Thus, survey descent displays desirable performance even on an objective not captured by the theory in sections 2–4.*

maximum of smooth functions, suggest that survey descent may retain this linear convergence behavior even on more general nonsmooth functions, hinting at its potential as a local nonsmooth, minimization technique.

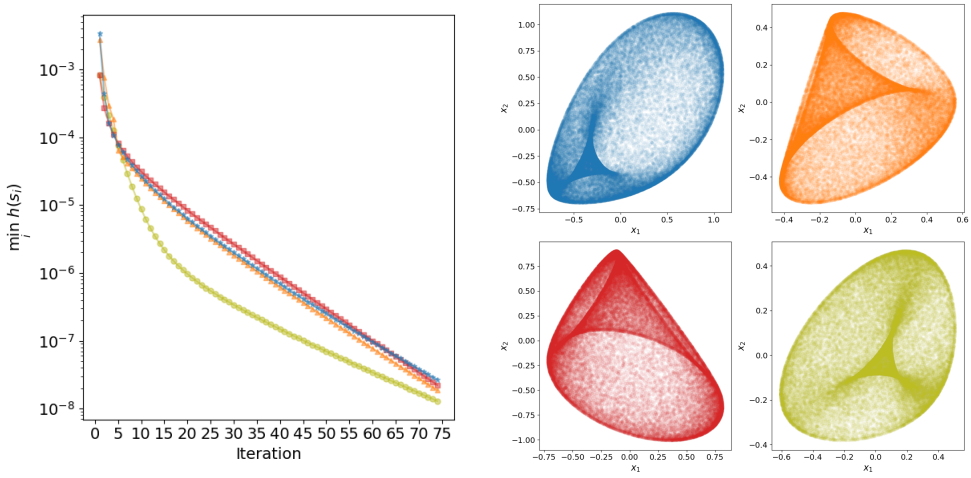
Many popular first-order methods exhibit empirical linear convergence on nonsmooth objectives. For example, our Figures 5 and 6 illustrate the linear convergence of BFGS when minimizing two simple nonsmooth objectives, and [31, 32] explicitly discuss this aspect of the algorithm. Similarly, the experiments of [47] show that first-order multipoint methods also display linear convergence when applied to a variety of nonsmooth machine learning tasks.

However, these behaviors are thought-provoking because modern nonsmooth, convex optimization *theory* has typically only proven that canonical first-order methods converge *sublinearly* when smoothness assumptions are absent—even in the presence of desirable properties such as strongly convex objectives: For example, see [38, Chapter 3.2] or [2, Chapter 8] that analyzes the subgradient method as well as [22] that analyzes mirror descent. Even for popular multipoint methods designed for nonsmooth optimization, previous theoretical guarantees have remained generally sublinear (discussed later in section 1.4). For comparison, in smooth settings, GD variants possess well-recognized *linear* convergence guarantees of on L-smooth and δ-strongly convex objectives [2, Theorem 10.29]. Our derivation of local linear convergence for survey descent on nonsmooth objectives will directly connect to these smooth GD results.

Survey Descent on Larger Eigenvalue Optimization Problems

$$h_{\mathrm{ME2}}(x) = \mathrm{MaxEigenvalue}\left(\begin{bmatrix} 0_{3\times 3} & 0_{3\times 7} \\ 0_{7\times 3} & -I_{7\times 7} \end{bmatrix} + \textstyle\sum_1^{40} x_i A_i\right) \quad x \in \mathbf{R}^{40}$$

for random symmetric $A_i \in \mathbf{R}^{10\times 10}$

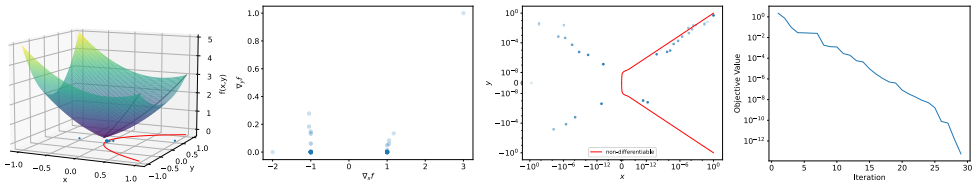


(a) Objective Value (b) Subdifferential Visualization

FIG. 4. *Survey descent on larger eigenvalue optimization problems. We use survey descent to minimize four instances of* h_{ME2}. *The set* $\{A_i\} = \{C_i^T + C_i\}$ *was generated from square matrices* C_i *with standard Gaussian entries, shifted to ensure linear independence and an optimal solution at zero. The initialization heuristic always chose* $\dim(\partial h_{\mathrm{ME2}}(0))+1 = 6$ *as the survey size without a priori information. Panel* (a) *plots best objective values over* 75 *iterations. Panel* (b) *shows two-dimensional projections of* $\partial h_{\mathrm{ME2}}(0)$ *formed by sampling* 10^5 *gradients within a ball of radius* 10^{-8}. *The outputs do not cluster around finitely many points, confirming heuristically that the instances are not max-of-smooth.*

Simple Max-Function Example with BFGS

$$h_{\max}(x, y) = \left|x - y^2\right| + x^2 + 2y^2$$



(a) (b) (c) (d)

FIG. 5. *Differentiability and linear convergence on nonsmooth objectives. We use BFGS to minimize the nonsmooth objective* $h_{\max}(x, y)$ *that achieves minimum value* 0 *at unique minimizer* $\bar{x} = (0, 0)$. *BFGS is implemented exactly as described in* [31], *initialized at* $(1, 0.5)$ *and ran for* 30 *iterations. Panel* (a) *visualizes* $h_{\max}$. *It shows the iterates (blue dots) in the xy-plane as well as the* $x = y^2$ *curve (red) on which* $h_{\max}$ *is nondifferentiable. Panels* (b) *and* (c), *respectively, show the gradients at and locations of each iterate with darker colors indicating later iterations. Panel* (b) *shows gradients form a convex hull that empirically resembles* $\partial h_{\max}(0, 0)$*'s one-dimensional structure. Panel* (c) *(in symmetric-log scale) shows that iterates lie within smooth subregions of the objective. Panel* (d) *records the linear convergence of the objective value.*

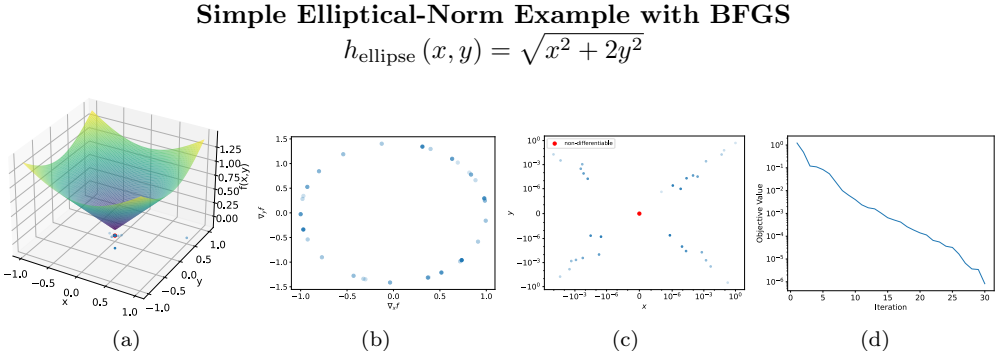


FIG. 6. *Differentiability and linear convergence on nonsmooth objectives. We use BFGS to minimize the nonsmooth objective* $h_{\text{ellipse}}(x, y)$ *that achieves minimum value* 0 *at unique minimizer* $\bar{x} = (0, 0)$. *BFGS is implemented exactly as described in* [31], *initialized at* $(1, 0.5)$ *and ran for* 30 *iterations. Panel* (a) *visualizes* h_{ellipse}. *It shows the iterates (blue dots) in the xy-plane as well as the origin (red) at which* $h_{\max}$ *is nondifferentiable. Panels* (b) *and* (c), *respectively, show the gradients at and locations of each iterate with darker colors indicating later iterations. Panel* (b) *shows gradients forming a convex hull that empirically resembles* $\partial h_{\text{ellipse}}(0, 0)$*'s two-dimensional structure. Panel* (c) *(in symmetric-log scale) shows all iterates lie within differentiable regions of the domain. Panel* (d) *records the linear convergence of the objective value.*

1.3. Motivation of survey descent. Before the formal analysis, we first describe and motivate some characteristic features within survey descent's design. First, note that the subproblems are simple quadratic programs with Euclidean ball constraints. Thus, they are efficiently solvable using second-order conic solvers. When all constraints moreover hold with equality, as occurs in our local convergence analysis, solving them further simplifies to routine linear algebra.

Next, the differentiability of input survey points is inspired by the common "differentiability of every iterate" behavior exhibited by a variety of popular first-order methods[1] when applied to nonsmooth objectives. To illustrate, Figures 5(c) and 6(c) show that the BFGS iterates minimizing the figures' nonsmooth objectives are always differentiable. This behavior also manifests in [31], for example, when minimizing the nonsmooth Rosenbrock and Chebyshev–Rosenbrock objectives. Recent works have begun developing theoretical explanations for this occurrence. In particular, [3] proved that, under standard assumptions, (stochastic) GD iterates are differentiable with probability one[2] even when objectives are nonconvex. Thus, we can easily imagine creating a survey of points at which h is differentiable by running BFGS or GD for a few initializing iterates and collecting some subset of those iterates. (For further discussion of implementation, see section 5.)

Lastly, the size k of the survey itself relates to the dimension of the subdifferential of h at its global minimizer $\bar{x}$ (assuming it exists), while the individual points $\{s_i\}_{i=1}^k$ "survey the landscape" of h near $\bar{x}$. The next two subsections elaborate on these intuitions.

[1] Bundle methods (discussed in section 1.4) are a notable exception. Their iterates often land exactly on the nonsmooth points of their associated objectives as a consequence of minimizing their underlying cutting-plane models.

[2] We distinguish the notion of "almost all iterates of a first-order method are differentiable" from the also common notion of "almost all points in the domain are differentiable"—describing how all locally Lipschitz functions are nondifferentiable on an at most measure-zero set in their domain. The latter result is often called Rademacher's theorem.

1.3.1. Dimension of the subdifferential. The dimensionality of subdifferentials captures key nonsmoothness properties. For example, smooth functions possess everywhere zero-dimensional subdifferentials since they contain only one point: the gradient. In comparison, Figure 5 shows an objective possessing a one-dimensional subdifferential at its minimizer characterized by

$$\partial h_{\max}(\bar{x}) = \operatorname{conv}\left(\left\{\begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}-1\\0\end{bmatrix}\right\}\right), \tag{1.5}$$

where $\operatorname{conv}(\cdot)$ denotes the convex hull. The objective is actually "partly smooth": It has a "smooth edge" in the $[0,1]^T$-direction orthogonal to $\partial h_{\max}(\bar{x})$. Partial smoothness is formalized and explored in detail by [30] and follow-up works.

Similarly, Figure 6 shows an objective where the subdifferential at its minimizer is a two-dimensional ellipse in the $\mathbf{R}^2$ domain:

$$\partial h_{\text{ellipse}}(\bar{x}) = \left\{(x,y) : \; x^2 + \frac{1}{2}y^2 \le 1\right\}. \tag{1.6}$$

The "full-dimensionality" of $\partial h_{\text{ellipse}}(\bar{x})$ captures the fact that h_{ellipse} is nondifferentiable in all $\mathbf{R}^2$ directions.

Our focus on subdifferentials is also empirically motivated: When employing first-order methods to minimize some objective h, the gradients of iterates near the global minimizer $\bar{x}$ typically have a convex hull whose dimension empirically coincides with the dimension of the subdifferential $\partial h(\bar{x})$. Figures 5(a) and (b) and 6(a) and (b) demonstrate this behavior. Moreover, this phenomenon occurs despite none of the iterates actually achieving the global minimum (Figures 5(c) and 6(c)). This is unsurprising since $\partial h(\bar{x})$ is the convex hull of the limit of gradients at nearby points [9, Theorem 2.5.1]. This motivates the creation of a "survey" of points near $\bar{x}$ such that their gradients capture the dimensionality of $\partial h(\bar{x})$. Introducing the max-of-smooth model clarifies this intuition.

1.3.2. Max-of-smooth model. As is well known, Fenchel conjugacy allows us to represent any continuous, convex function as the supremum of a family of affine functions (see, for example, [6, Chapter 3.3]). More generally, continuous, convex functions are instances of *lower-$\mathcal{C}^1$ functions*, which are functions representable as maximums of smooth and compactly parameterized functions.

Definition 2 (lower-$\mathcal{C}^1$ functions; [46], [10, Corollary 3]). *A locally Lipschitz function $h : \mathbf{R}^n \to \mathbf{R}$ is called lower-$\mathcal{C}^1$ if, for every $\bar{x} \in \mathbf{R}^n$, there exists a compact parameterizing set $\boldsymbol{T}$, a neighborhood $\boldsymbol{X}$ around $\bar{x}$, and functions $g : \boldsymbol{T} \times \boldsymbol{X} \to \mathbf{R}$ such that $g(t,x)$ and $\nabla_x g(t,x)$ are jointly continuous in t and x and if*

$$h(x) = \max_{t \in \boldsymbol{T}} g(t,x). \tag{1.7}$$

Lower-$\mathcal{C}^1$ functions are exactly the locally Lipschitz and approximately convex functions, where "approximately convex" relaxes the canonical notion of convexity and is defined in [10].

Lower-$\mathcal{C}^1$ functions subsume the entire class of continuous, convex objectives. Definition 2 implies that, for any global minimizer $\bar{x}$ of the objective h, there exists a family of functions $\{g(t,\cdot)\}_{t \in \boldsymbol{T}}$ satisfying (1.7) at $x = \bar{x}$. This family also allows us to describe h's subdifferential at its minimizer:

$$\partial h(\bar{x}) = \operatorname{conv}\left\{\nabla_x g(t,\bar{x}) : \; t \in \boldsymbol{T}'\right\}, \tag{1.8}$$

where $\boldsymbol{T}' = \{t : \ g(t, \bar{x}) = h(\bar{x})\}$. For more details on (1.8), see [44, Theorem 10.31]. This characterization inspires us to consider a structurally revealing approximating model for convex, nonsmooth objectives where $\boldsymbol{T}$ is *finite* and $\boldsymbol{T} = \boldsymbol{T}'$. In this setting, (1.7) reduces to a *max-of-smooth* function:

$$f(x) = \max_{i=1,\dots,k} f_i(x), \tag{1.9}$$

where k is a finite integer and each f_i is a differentiable function. Just like in classical analyses of GD (see, for example, [2, Chapter 10.6]), we will assume that each f_i is a $\mathcal{C}^2$ function and strongly convex.

When the objective is a max-of-smooth function (1.9), we could "survey" the landscape of f by somehow obtaining a point, s_i, from each region $\mathcal{R}_i \equiv \{x : f_i(x) > f_j(x) \ \forall j \neq i\}$ for all i. If the gradients $\{\nabla f_i(\bar{x})\}_{i=1}^k$ are moreover affinely independent—meaning $\bar{x}$ is a "nondegenerate" minimizer—and the s_i's are sufficiently close to $\bar{x}$, the dimension of the convex hull of the survey gradients, $\text{conv}(\{\nabla f_i(s_i)\}_{i=1}^k)$, will coincide with the dimension of the objective's subdifferential at its minimizer, $\partial f(\bar{x}) = \text{conv}(\{\nabla f_i(\bar{x})\}_{i=1}^k)$. Moreover, the dimension of $\partial f(\bar{x})$ is then $k-1$. In other words, it is exactly one less than the size of the survey.

Next, observe from Definition 1 that each survey descent subproblem simply performs one projected GD step onto the feasible region defined by (1.4). The max-of-smooth objective intuitively motivates this region. In this case, when survey descent is equipped with an initializing survey consisting of one point in each $\mathcal{R}_i$ as discussed above, $f(s_i) = f_i(s_i)$ and $\nabla f(s_i) = \nabla f_i(s_i)$ for all i. Then, since the components f_i are L-smooth, the constraints for the ith survey descent subproblem (1.4) restrict solutions to a region where the linear lower bound of f_i is at least the quadratic upper bounds of the remaining f_j, $j \neq i$. Therefore, when the ith subproblem is feasible, its output s_i^+ would necessarily remain within $\mathcal{R}_i$. Section 2 formalizes these ideas in detail, but their geometry is simple: Figure 1 shows the behavior of survey descent on a simple max-of-smooth objective on $\mathbf{R}^2$, while Figure 7 presents an abstract illustration of the above intuitions.

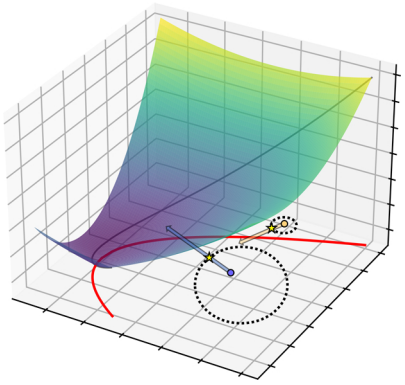


FIG. 7. *Intuition of survey descent. Abstract depiction of survey descent on max-of-smooth objectives* (1.9) *with* $k = 2$ *components and using two survey points (circular dots). Each survey point,* s_i*, is associated with one* f_i*-component of the objective for* $i \in \{1, 2\}$*. In the* i*th survey descent subproblem, a gradient step (arrow) from the* i*th survey point is projected onto the subproblem's constraints (illustrated with dashed-line boundaries). The constraints* (1.4) *prevent subproblem outputs (stars) from crossing the "nonsmooth boundary" (red line). In other words, the* i*th output must remain in the subregion of the domain where the* f_i*-component is "active."*

Despite our discussions of $\{f_i\}_{i=1}^k$, note the very important feature that *performing the survey descent iteration on max-of-smooth function objectives does not require access to individual f_i's*. It only involves function value and gradient evaluations of the overall objective f. Moreover—although survey descent is intuitively motivated by the max-of-smooth function objective and possesses a structurally revealing local linear convergence theory on these objectives (sections 2–4)—it is broadly implementable on *any* objective function. Indeed, survey descent still displays promising behavior even on objectives *not* expressible as the maximum of finitely many smooth functions. For instance, Figures 3 and 4 show the visibly linear convergence of survey descent on max-matrix-eigenvalue objectives whose eigenvalues are nonsmooth functions of the input variables.

1.4. Related works. A well-known example of [37] shows sublinear worst-case global complexity for any first-order method for arbitrary nonsmooth convex objectives. That example focuses on an initial sequence of iterates of length less than the dimension of the domain. In contrast, we prove that survey descent achieves linear convergence, but only locally and asymptotically, and on an interesting subclass of objectives.

Also worth noting for comparison is classical work on minimizing the maximum of finitely many *given* smooth functions f_i, where one can access each component f_i rather than just their pointwise maximum, as in survey descent. Such problems are standard in classical optimization, easily solved by reduction to classical nonlinear programs. [24] developed a more sophisticated two-phase method, first identifying a search direction by solving a quadratic program built from the function values f_i and gradients ∇f_i and then performing a line search. More recently, [38, Scheme 2.3.13] computes a linearization of each f_i-component, $\{\ell_{\tilde{x}}^{f_i}(x)\}_{i=1}^k$, at some iterate $\tilde{x}$; performs a proximal-point iteration on $\max_i \ell_{\tilde{x}}^{f_i}(x)$; and then determines the next iteration using a carefully designed momentum step. Both [24] and [38, Scheme 2.3.13] achieve global, linear convergence on strongly convex, max-of-smooth function objectives.

At large scale, motivated by modern machine learning applications, [8] describes a more sophisticated approach. Using function value and gradient evaluations of the f_i-components, [8] implements a ball regularized optimization oracle (BROO) that returns the minimizer of the objective—subject to a proximal-regularizer—within a ball around a queried point. Using this BROO, [8] then designs an efficient minimization procedure by repeatedly updating an iterate using subroutines of line searches and BROO calls. [8] prove global, sublinear convergence rates for this procedure on max-of-smooth function objectives.

Within the derivative-free setting, researchers have also studied max-of-smooth objectives [21] as well as generalizations into (potentially nonconvex) objectives that are the composition of multiple component functions [23]. By leveraging careful line searches and (approximations to) active set information at individual iterates, these methods guarantee convergence using only function-value evaluations of individual objective components without any need for gradient evaluations. Convergence rates are challenging to derive in derivative-free settings and are generally sublinear where they do exist (for example, in [19]). For more detailed discussions on works investigating such derivative-free methods, see [23].

[24, 38, 8] and the derivative-free methods above differ from survey descent since they crucially rely on access to the individual components of a max function objective. In comparison, survey descent only assumes a first-order oracle that returns the

function value and gradient of f *without* access to any component f_i's. Thus, survey descent is implementable on *any* objective, even those without max-of-smooth structure, suggesting a promising future research direction. Also notable is the distinction in computational style: survey descent subproblems consist of k parallel projected GD steps, updating an entire collection of survey points rather than only one iterate as in the works above.

On the other hand, more sophisticated structural oracles can lead even to superlinear convergence on nonsmooth objectives, as remarked in several prior works. For eigenvalue optimization, [42] and [40] develop quadratically convergent procedures by incorporating in particular a Hessian oracle within a nearby $\mathcal{U}$-subspace [29]. On general nonsmooth objectives, using $\mathcal{U}$-Hessians and proximal operators on the objective, [35] presents a $\mathcal{VU}$-algorithm whose serious steps converge superlinearly. In contrast, although survey descent conceptually relies on a max-of-smooth model, its implementation is general, requiring only function and gradient evaluations.

Among multipoint methods, one popular class for nonsmooth optimization is bundle methods [27, 49], which possess a long, successful history (described in [36]). In its most transparent proximal form, they rely on a multipoint collection (a "bundle") to build a piecewise-linear, cutting-plane model that is minimized in each step with a proximal operator around a current iterate (a "center"). If the proximal-step outputs a point that sufficiently decreases the function value, the method takes a "serious step" updating the center with the outputted point; otherwise, the method takes a "null step" which does not update the center but adds the outputted point to the running point collection to improve the cutting-plane model.

The related "level-bundle" approach [14, 25, 28] also shares some similarities with survey descent. Each iteration projects the current center onto a sublevel set for the current cutting-plane model, the update being either accepted if the objective value decreases satisfactorily—a serious step—or rejected, in which case the cutting-plane model is updated. The projection ingredient is similar in both methods, but survey descent solves k *parallel* subproblems at each iteration, whereas level-bundle methods perform projections sequentially, enforcing objective decrease for serious steps.

Bundle methods work well in practice [5, 18, 45, 48] and enjoy a robust global convergence theory: see [13] for a comprehensive survey. In particular, relative to the number of *serious steps*, bundle methods converge linearly on convex objectives whose subdifferentials satisfy certain growth conditions away from the minimizer, as a consequence of a more general framework of [43] and developed further in [1].[3]

However, relative to both null and serious steps, prior published studies [15, 16, 26] have only derived sublinear convergence guarantees, even when the objective is strongly convex (although [12] is a recent promising advance). Two recent works [33, 34] unify and present optimal iteration complexities for proximal bundle methods, all of which converge sublinearly, even on strongly convex objectives. In contrast, we show in this paper that the survey descent iteration—at least in the case of strongly convex, max-of-smooth function objectives—achieves a *local*, *linear* convergence rate.

[3]Two independent related works [1, 12] were announced recently. The first is a flexible, unified analysis of linearly convergent descent methods on weakly convex objectives, in particular covering the serious step sequence for bundle methods. A transparent analysis of survey descent in this framework is not immediately apparent; the original approach we present here is direct. The second announcement presents an interesting blackbox randomized first-order method that is nearly linear convergent with high probability.

Finally, we remark that survey descent surveys and bundle method bundles serve different functional roles. In bundle methods, null steps query points sequentially to improve the objective model around the center. In survey descent, there are neither auxiliary null steps nor a center. Instead, a fixed-size survey of points is cautiously updated *in parallel* trying to mimic the steady progress of GD.

2. Survey descent on the max-of-smooth objectives. We are particularly interested in convex objective functions f with a *nondegenerate minimizer* $\bar{x}$:

$$\bar{x} = \arg\min_{x\in\mathbf{R}^n} f(x) \text{ and } 0 \in \operatorname{relint}\left(\partial f(\bar{x})\right), \tag{2.1}$$

where $\operatorname{relint}(\cdot)$ denotes the relative interior. We further assume the following structure around $\bar{x}$.

DEFINITION 3. *A function $f:\mathbf{R}^n\to\mathbf{R}$ is a strong $\mathcal{C}^2$ max function if it is locally expressible near $\bar{x}$ as*

$$f(x) = \max_{i=1,\dots k} f_i(x), \tag{2.2}$$

where k is some finite number, the components $\{f_i\}_{i=1}^k$ are $\mathcal{C}^2$-functions satisfying $f_i(\bar{x}) = f(\bar{x})$ for all i, their gradients $\{\nabla f_i(\bar{x})\}_{i=1}^k$ are affinely independent, and their Hessians $\nabla^2 f_i(\bar{x})$ are positive definite. As a consequence, there exists constants $\delta, L>0$ such that their Hessians satisfy

$$\delta I \preceq \nabla^2 f_i(x) \preceq LI \ \forall\ i = 1,\dots,k, \tag{2.3}$$

for x near $\bar{x}$—where I is the $n\times n$ identity matrix.

Our analysis in this paper is entirely local, but we assume for the rest of this paper that f is a strong $\mathcal{C}^2$ max function with nondegenerate minimizer $\bar{x}$. For simplicity, we also assume that the properties in Definition 3 hold *globally* for all $x\in\mathbf{R}^n$. In this setting, we refer to k—which is necessarily unique—as the *degree* of f. Moreover, note that (2.3) implies all components $\{f_i\}_{i=1}^k$ are L-smooth, both f and its components are δ-strongly convex, and $\bar{x}$ is f's *unique* minimizer.

To study survey descent (Definition 1) on the objective f, we choose the above-presented L as the step-control parameter and define the following notions of *valid* and *minimizing* surveys.

DEFINITION 4 ($\mathcal{S},\bar{\mathcal{X}}$-surveys and validity). *For the objective f, define a survey as a matrix $\mathcal{S}\equiv[s_1,\dots,s_k]\in\mathbb{R}^{n\times k}$ consisting of k columns[4] (the "survey points"), where k is the degree of f; valid surveys as surveys satisfying $f_i(s_i) > f_j(s_i)$ for all $i\neq j$; and the minimizing survey as $\bar{\mathcal{X}}\equiv[\bar{x},\dots,\bar{x}]$ whose survey points are all equal to $\bar{x}$.*

On the space of surveys, we adopt the norm

$$\|\mathcal{S}\| \equiv \|[s_1,\dots,s_k]\|_{2,\infty} = \max_{i=1,\dots,k} \|s_i\|_2, \tag{2.4}$$

[4]We can equivalently consider $\mathcal{S}$ as a set $\{s_i\}_{i=1}^k$. We will use the matrix and set interpretations interchangeably.

which allows us to quantify the distance between a survey $\mathcal{S}$ from $\bar{\mathcal{X}}$ using $\|\mathcal{S}-\bar{\mathcal{X}}\|$, an intuitive measure of the distance to the furthest survey point. The precise choice of norm is immaterial to our computations and theory.

Given a valid $\mathcal{S}$ sufficiently close to $\bar{\mathcal{X}}$, survey descent exhibits many desirable properties when minimizing f. As a preliminary, note that (2.1) and Definition 3 imply there exists unique *critical weights* $\{\bar{\lambda}_i\}_{i=1}^k > 0$ such that

$$\sum_{i=1}^{k} \bar{\lambda}_i = 1, \tag{2.5}$$

$$\sum_{i=1}^{k} \bar{\lambda}_i \nabla f_i(\bar{x}) = 0. \tag{2.6}$$

As we will see, the Lagrange multipliers of survey descent subproblems will lie "close" to $\{\bar{\lambda}_i\}_{i=1}^k > 0$, while the subproblem outputs $\{s_i^+\}_{i=1}^k$ will lie near $\bar{x}$. To characterize the order of magnitude of these distances, we adopt the "Big-Oh" notation. In particular, for a mapping $g : \mathbf{E} \to \mathbf{F}$ between two Euclidean spaces and letting $\|\cdot\|^p$ denote an arbitrary Euclidean norm raised to the pth power, we use the notation $g(x) = O(\|x\|^p)$ to indicate the property that there exists a constant $K > 0$ such that $\|g(x)\| \leq K\|x\|^p$ holds for all small x. By itself, we let $\mathrm{O}(\|x\|^p)$ denote an element of the class of all functions with this property.

With this terminology, we present our primary theoretical setting as well as our first result.

SETTING A (local analysis of survey descent). *Consider an iteration of survey descent on a strong $\mathcal{C}^2$ max function objective, f, with a survey $\mathcal{S}$ that is valid and sufficiently close to $\bar{\mathcal{X}}$. For $i = 1, \ldots, k$, denote the output of the ith survey descent subproblem by s_i^+.*

THEOREM 5 (local feasibility, uniqueness, tightness, and smoothness of survey descent). *Assume Setting A. Then for all i, the survey descent subproblem $(P_i^{\mathcal{S}})$ is feasible and has a unique solution s_i^+, which satisfies all the constraints with equality, with unique associated Lagrange multipliers λ_j^i for the jth constraint within the ith subproblem (for $j \neq i$). The solutions and multipliers depend smoothly on the input survey $\mathcal{S}$ and satisfy*

$$s_i^+(\mathcal{S}) = \bar{x} + \mathrm{O}\left(\left\|\mathcal{S} - \bar{\mathcal{X}}\right\|\right) \qquad \text{and} \qquad \lambda_j^i(\mathcal{S}) = \bar{\lambda}_j + \mathrm{O}\left(\left\|\mathcal{S} - \bar{\mathcal{X}}\right\|\right) \ \forall\ j \neq i. \tag{2.7}$$

Proof. The proof follows from a routine analysis of the subproblem's first-order optimality conditions combined with the implicit function theorem. ∎

We refer to Appendix B of [20] for a more direct and elementary proof that is computationally revealing. In particular, it shows we can compute $s_i^+(\mathcal{S})$ and $\lambda_j^i(\mathcal{S})$ entirely with linear algebra routines and a scalar square-root.

Theorem 5 plays a key role in deducing many theoretical results in this paper. Most immediately, under the assumptions of the theorem, the output survey $\mathcal{S}^+ = \{s_i^+\}_{i=1}^k$ of a survey descent iteration will also be valid. First, note that f is differentiable at every valid survey point and the associated gradients satisfy

$$\nabla f(s_i) = \nabla f_i(s_i) \ \forall\ i = 1, \ldots, k, \tag{2.8}$$

which leads to the following observation.

OBSERVATION 6 (survey descent iterations on max-of-smooth objectives). *Assume that $\mathcal{S}$ is a valid survey for the objective f. Then, for all i, the ith subproblem $(P_i^{\mathcal{S}})$ of survey descent (Definition 1) is equivalent to*

$$\begin{aligned} &\min_x \ \left\| x - \left(s_i - \frac{1}{L}\nabla f_i(s_i) \right) \right\|_2^2 \\ &\text{s.t. } \ell_{s_j}^{f_j}(x) + \frac{L}{2}\|x - s_j\|_2^2 \le \ell_{s_i}^{f_i}(x) \ \forall\ j \neq i. \end{aligned} \tag{2.9}$$

Using Theorem 5 and Observation 6, we deduce the aforementioned validity of $\mathcal{S}^+$.

THEOREM 7 (preservation of validity). *Assuming Setting A, the output survey $\mathcal{S}^+$ of a survey descent iteration is valid.*

Proof. Consider any fixed i. If $s_i^+ = s_i$, then the fact that $f_i(s_i^+) > f_j(s_i^+)$ for all $j \neq i$ immediately follows since Theorem 5 assumes $\mathcal{S}$ is valid.

Now, consider the $s_i^+ \neq s_i$ case. By Theorem 5, s_i^+ satisfies the constraints of $(P_i^{\mathcal{S}})$ with equality. Using Observation 6, we express this as

$$\ell_{s_i}^{f_i}(s_i^+) = \ell_{s_j}^{f_j}(s_i^+) + \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 \ \forall\ j \neq i. \tag{2.10}$$

The δ-strong convexity of f_i implies

$$f_i\left(s_i^+\right) \ge \ell_{s_i}^{f_i}(s_i^+) + \frac{\delta}{2}\left\|s_i^+ - s_i\right\|_2^2 > \ell_{s_i}^{f_i}(s_i^+),$$

where the strict inequality follows from $s_i^+ \neq s_i$. Next, by the L-smoothness of $\{f_j\}_{j=1}^k$, the right-hand side of (2.10) is a s_j-centered quadratic upper bound of f_j evaluated at s_i^+. Thus,

$$\ell_{s_j}^{f_j}(s_i^+) + \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 \ge f_j\left(s_i^+\right) \ \forall\ j \neq i.$$

Therefore,

$$f_i\left(s_i^+\right) > \ell_{s_i}^{f_i}(s_i^+) = \ell_{s_j}^{f_j}(s_i^+) + \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 \ge f_j\left(s_i^+\right) \ \forall\ j \neq i.$$

This completes the proof. ∎

3. Connecting survey descent and GD. We will build a local convergence theory by connecting survey descent to projected GD steps on the smooth components $\{f_i\}_{i=1}^k$ of the objective f. To form this connection, we first identify the affine $\overline{\mathcal{U}}$-*subspace* possessing the following equivalent characterizations:

$$\overline{\mathcal{U}} \equiv \bar{x} + \operatorname{span}\{\nabla f_i(\bar{x}) - \nabla f_j(\bar{x}) : 1 \le i, j \le k\}^{\perp} \tag{3.1}$$

$$= \bar{x} + \left\{ \sum_i^k \gamma_i \nabla f_i(\bar{x}) : \sum_i^k \gamma_i = 0 \right\}^{\perp} \tag{3.2}$$

$$= \bar{x} + \left\{ x \in \mathbf{R}^n : \ \nabla f_i(\bar{x})^T x = \nabla f_j(\bar{x})^T x \ \forall\ 1 \le i, j \le k \right\}. \tag{3.3}$$

$\overline{\mathcal{U}}$ is so named because the $\bar{x}$-centered, linear subspace $\{\overline{\mathcal{U}} - \bar{x}\}$ is commonly called the "$\mathcal{U}$-subspace" [29] and captures the directions in which f is smooth around $\bar{x}$. Theorem 11 will show that the outputs of survey descent subproblems are approximately—up to an $\mathrm{O}(\|\mathcal{S} - \bar{\mathcal{X}}\|^2)$ term—a convex combination of $\overline{\mathcal{U}}$-projected GD steps on the smooth components $\{f_i\}_{i=1}^k$.

To prove Lemmas 8–10 leading to Theorem 11, we adopt the following notation for the gradients and Hessians of the components of f (Definition 3):

$$a_i \equiv \nabla f_i(\bar{x}) \text{ and } A_i \equiv \nabla^2 f_i(\bar{x}) \ \forall \ i = 1, \dots, k. \tag{3.4}$$

Additionally, assuming $\bar{x} = 0$ without loss of generality will simplify many of our proofs. In this case, the second-order Taylor expansion of f_i around zero is

$$f_i(x) = f(0) + a_i^T x + \frac{1}{2} x^T A_i x + r_i(x) \ \forall \ i = 1, \dots, k, \tag{3.5}$$

where $r_i(x)$ is a residual function.

LEMMA 8. *For all i, the residual function $r_i(x)$ in (3.5) satisfies $r_i(x) = \mathrm{O}(\|x\|^2)$ and $\nabla r_i(x) = \mathrm{O}(\|x\|)$.*

Proof. $r_i(x) = \mathrm{O}(\|x\|^2)$ follows directly from Taylor's theorem. Next, observe that

$$r_i(x) = f_i(x) - \left(f(0) + a_i^T x + \frac{1}{2} x^T A_i x \right).$$

Thus, since f_i is $\mathcal{C}^2$, the residual r_i must also be $\mathcal{C}^2$, so its gradient and Hessian are well defined. This implies $\nabla r_i(x)$ is then $\mathcal{C}^1$. Taylor expanding ∇r_i around zero then gives the desired $\nabla r_i(x) = \mathrm{O}(\|x\|)$ after observing that $\nabla r_i(0)$ is the zero-vector and $\nabla^2 r_i(0)$ is the zero-matrix (since r_i is the residual of the Taylor expansion of f_i around zero). ∎

We now prove and present the main results of this section.

LEMMA 9. *Assume Setting A, and suppose $\bar{x} = 0$. Then, s_i^+ satisfies*

$$\begin{aligned} \mathcal{P}_{\overline{\mathcal{U}}}\left[s_i^+\right] = \frac{1}{2 - \bar{\lambda}_i} \mathcal{P}_{\overline{\mathcal{U}}}\left[s_i - \frac{\bar{\lambda}_i}{L} \nabla f_i\left(s_i\right) \right] \\ + \sum_{j \neq i} \frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i} \mathcal{P}_{\overline{\mathcal{U}}}\left[s_j - \frac{1}{L} \nabla f_j(s_j) \right] + \mathrm{O}\left(\|\mathcal{S}\|^2 \right) \ \forall \ i = 1, \dots, k, \end{aligned}$$

where $\mathcal{P}_{\overline{\mathcal{U}}}$ is the projection operator onto the $\overline{\mathcal{U}}$-subspace (3.1).

Proof. Without loss of generality, assume $i = k$. Scale the objective of problem (2.9) by $\frac{L}{2}$, and consider the equivalent problem

$$\begin{aligned} \min_x \ & \frac{L}{2} \left\| x - \left(s_k - \frac{1}{L} \nabla f_k\left(s_k\right) \right) \right\|_2^2 \\ \text{s.t. } & \ell_{s_j}^{f_j}(x) + \frac{L}{2} \left\| x - s_j \right\|_2^2 \leq \ell_{s_k}^{f_k}(x) \ \forall \ j \neq k. \end{aligned} \tag{3.6}$$

By Theorem 5 and Observation 6, there exist unique Lagrange multipliers $\{\lambda_j\}_{j=1}^{k-1}$ (we omit the k-superscript for simplicity) satisfying the following first-order optimality condition for (3.6):

$$0 = L\left(s_k^+ - \left(s_k - \frac{1}{L} \nabla f_k\left(s_k\right) \right) \right) + \sum_{j=1}^{k-1} \lambda_j \left(\nabla f_j\left(s_j\right) + L\left(s_k^+ - s_j\right) - \nabla f_k\left(s_k\right) \right).$$

After rearranging this gives

$$0 = L\left(1+\sum_{j=1}^{k-1}\lambda_j\right)s_k^+ - L\left(s_k+\sum_{j=1}^{k-1}\lambda_j s_j\right)+\nabla f_k(s_k)+\sum_{j=1}^{k-1}\lambda_j\left(\nabla f_j(s_j)-\nabla f_k(s_k)\right). \tag{3.7}$$

Define the following scalars:

$$\lambda_k \equiv 1-\sum_{j=1}^{k-1}\lambda_j, \tag{3.8}$$

$$\mu_j \equiv \bar{\lambda}_j - \lambda_j \;\forall\; j=1,\ldots,k. \tag{3.9}$$

For $j \neq k$, Theorem 5 gives $\mu_j = \mathrm{O}(\|\mathcal{S}\|)$. For $j=k$, combining (3.8)–(3.9) with (2.5) implies

$$\mu_k = \bar{\lambda}_k - \left(1-\sum_{j=1}^{k-1}\lambda_j\right) = \bar{\lambda}_k - \left(1-\sum_{j=1}^{k-1}\bar{\lambda}_j\right) + \mathrm{O}\left(\|\mathcal{S}\|\right) = \mathrm{O}\left(\|\mathcal{S}\|\right).$$

Thus,

$$\mu_j = \mathrm{O}(\|\mathcal{S}\|) \;\forall\; j=1,\ldots,k. \tag{3.10}$$

Substituting (3.9) into (3.7), applying identities (2.5)–(2.6), collecting $\mathrm{O}(\|\mathcal{S}\|^2)$ terms, and solving for s_k^+ gives

$$\begin{aligned} s_k^+ = {} & \frac{1}{2-\bar{\lambda}_k}\left[s_k - \frac{\bar{\lambda}_k}{L}\nabla f_k\left(s_k\right)\right] + \sum_{j=1}^{k-1}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_k}\left[s_j - \frac{1}{L}\nabla f_j(s_j)\right] \\ & + \frac{1}{L\left(2-\bar{\lambda}_k\right)}\sum_{j=1}^{k}\mu_j\nabla f_j(s_j) + \mathrm{O}\left(\|\mathcal{S}\|^2\right). \end{aligned} \tag{3.11}$$

Adopting the notation (3.4) and applying the Taylor expansion (3.5), observe

$$\sum_{j=1}^{k}\mu_j\nabla f_j(s_j) = \sum_{j=1}^{k}\mu_j\left(a_j + A_j s_j + \nabla r_j(s_j)\right) = \sum_{j=1}^{k}\mu_j a_j + \mathrm{O}\left(\|\mathcal{S}\|^2\right), \tag{3.12}$$

where the last equality follows from (3.10) and Lemma 8. Substituting (3.12) into (3.11) then leads to

$$\begin{aligned} s_k^+ = {} & \frac{1}{2-\bar{\lambda}_k}\left[s_k - \frac{\bar{\lambda}_k}{L}\nabla f_k\left(s_k\right)\right] + \sum_{j=1}^{k-1}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_k}\left[s_j - \frac{1}{L}\nabla f_j(s_j)\right] \\ & + \frac{1}{L\left(2-\bar{\lambda}_k\right)}\underbrace{\sum_{j=1}^{k}\mu_j a_j}_{(\star)} + \mathrm{O}\left(\|\mathcal{S}\|^2\right). \end{aligned} \tag{3.13}$$

Next, again combine (3.8)–(3.9) with (2.5) to deduce

$$\sum_{j=1}^{k}\mu_j = \sum_{j=1}^{k}\left(\bar{\lambda}_j - \lambda_j\right) = 0.$$

Then, it follows from (3.2) that $\mathcal{P}_{\overline{\mathcal{U}}}[(\star)] = 0$. The desired result then follows from applying $\mathcal{P}_{\overline{\mathcal{U}}}$ to (3.13). ∎

LEMMA 10. *Assume Setting A, and let* $\bar{x} = 0$. *Then,* s_i^+ *satisfies*

$$\mathcal{P}_{\overline{\mathcal{U}}^\perp}\left[s_i^+\right] = \mathrm{O}\left(\|\mathcal{S}\|^2\right) \ \forall\ i = 1, \ldots, k,$$

where $\mathcal{P}_{\overline{\mathcal{U}}^\perp}$ *is the projection operator onto the orthogonal complement of the* $\overline{\mathcal{U}}$*-subspace* (3.1).

Proof. Consider any fixed i. By Theorem 5, the constraints of $(P_i^{\mathcal{S}})$ are tight. Thus, by Observation 6, s_i^+ satisfies

$$f_j(s_j) + \nabla f_j(s_j)^T\left(s_i^+ - s_j\right) + \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 = f_i(s_i) + \nabla f_i(s_i)^T\left(s_i^+ - s_i\right) \ \forall\ j \neq i.$$

Using (3.5) in the above gives, for all $j \neq i$,

$$\begin{aligned} &f(0) + a_j^T s_j + \frac{1}{2} s_j^T A_j s_j + r_j(s_j) + \left(a_j + A_j s_j + \nabla r_j(s_j)\right)^T\left(s_i^+ - s_j\right) + \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 \\ &\quad = f(0) + a_i^T s_i + \frac{1}{2} s_i^T A_i s_i + r_i(s_i) + \left(a_i + A_i s_i + \nabla r_i(s_i)\right)^T\left(s_i^+ - s_i\right). \end{aligned}$$

After simplification and rearrangement, we obtain

$$\begin{aligned} \left(a_j - a_i\right)^T s_i^+ = &-\left(A_j s_j - A_i s_i + \nabla r_j(s_j) - \nabla r_i(s_i)\right)^T s_i^+ \\ &- \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 - c_j(s_j) + c_i(s_i) \ \forall\ j \neq i, \end{aligned} \tag{3.14}$$

where we define the functions

$$c_j(x) \equiv -\frac{1}{2} x^T A_j x + r_j(x) - \nabla r_j(x)^T x \ \forall\ j = 1, \ldots, i.$$

By Lemma 8, $c_j(x)$ is $O(\|x\|^2)$. Since we assume $\bar{x} = 0$, by Theorem 5, $s_i^+ = \mathrm{O}(\|\mathcal{S}\|)$. We then deduce that all terms on the right-hand side of (3.14) are $\mathrm{O}(\|\mathcal{S}\|^2)$:

$$\left(a_j - a_i\right)^T s_i^+ = \mathrm{O}\left(\|\mathcal{S}\|^2\right) \ \forall\ j \neq i. \tag{3.15}$$

By (3.1),

$$\overline{\mathcal{U}}^\perp = \mathrm{span}\left\{a_i - a_j : 1 \leq i, j \leq k\right\} = \mathrm{span}\left\{a_j - a_i :\ j \neq i\right\}.$$

Combined with the affine independence of $\{a_j\}_{j=1}^k$ (Definition 3), the above implies $\{a_j - a_i\}_{j \neq i}$ forms a basis for $\overline{\mathcal{U}}^\perp$. Therefore, we can express the projection onto $\overline{\mathcal{U}}^\perp$ in terms $\{a_j - a_i\}_{j \neq i}$. The result now follows from (3.15) and standard linear algebra. ∎

THEOREM 11 (gradient descent approximation of survey updates). *Assume Setting A. Then the solution of the* i*th survey descent subproblem satisfies, for all* i,

$$s_i^+ = \tilde{s}_i + \mathrm{O}\left(\left\|\mathcal{S} - \bar{\mathcal{X}}\right\|^2\right), \tag{3.16}$$

where we define

$$\tilde{s}_i \equiv \frac{1}{2 - \bar{\lambda}_i}\mathcal{P}_{\overline{\mathcal{U}}}\left[s_i - \frac{\bar{\lambda}_i}{L}\nabla f_i(s_i)\right] + \sum_{j \neq i}\frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i}\mathcal{P}_{\overline{\mathcal{U}}}\left[s_j - \frac{1}{L}\nabla f_j(s_j)\right], \tag{3.17}$$

and $\mathcal{P}_{\overline{\mathcal{U}}}$ *is the projection operator onto the* $\overline{\mathcal{U}}$*-subspace* (3.1).

Proof. Without loss of generality, assume that $\bar{x} = 0$. Then, the desired result follows from decomposing $s_k^+ = \mathcal{P}_{\overline{\mathcal{U}}}[s_k^+] + \mathcal{P}_{\overline{\mathcal{U}}^\perp}[s_k^+]$ and applying Lemmas 9 and 10. We can deduce the general $\bar{x} \neq 0$ case with the change of variables $x \leftarrow (x - \bar{x})$. ∎

4. Local linear convergence of surveys. Observe that the constituent terms of $\tilde{s}_i$ in (3.17) are $\overline{\mathcal{U}}$-projected gradient steps on the L-smooth and δ-strongly convex component functions $\{f_i\}_{i=1}^k$ (Definition 3). Thus, we can apply classical theory on projected GD described in standard texts such as [2, Theorem 10.29]. The idea of tracking iterates relative to their projections onto some nearby "active" manifold is familiar in nonsmooth optimization. Recent examples include [7, 17, 11].

THEOREM 12 (projected GD behavior). *Assume that $\mathcal{S}$ is a valid survey for the strong $\mathcal{C}^2$ max function objective f. Then, the following holds for all i and any constant $\tilde{L} \geq L$:*

$$\left\| \mathcal{P}_{\overline{\mathcal{U}}} \left[s_i - \frac{1}{\tilde{L}} \nabla f_i(s_i) \right] - \bar{x} \right\|_2^2 \leq \left(1 - \frac{\delta}{\tilde{L}} \right) \|s_i - \bar{x}\|_2^2,$$

$$f_i \left(\mathcal{P}_{\overline{\mathcal{U}}} \left[s_i - \frac{1}{\tilde{L}} \nabla f_i(s_i) \right] \right) - f_i(\bar{x}) \leq \frac{\tilde{L}}{2} \left(1 - \frac{\delta}{\tilde{L}} \right) \|s_i - \bar{x}\|^2.$$

Proof. Consider any fixed i. It is easy to check that $\bar{x} = \arg\min_{x \in \overline{\mathcal{U}}} f_i(x)$ using first-order optimality conditions combined with (2.5)–(2.6) and (3.3). The desired result then follows from classical projected GD convergence results (given in standard texts such as [2, Theorem 10.29]) after observing that $\overline{\mathcal{U}}$ is convex and f_i is L-smooth and δ-strongly convex. ∎

To prove local linear convergence, we show below that the survey points contract towards $\bar{x}$. The corresponding contraction ratios will depend on the minimum critical weight defined as follows:

$$\bar{\lambda}_{\min} \equiv \min_{i=1,\ldots,k} \bar{\lambda}_i > 0. \tag{4.1}$$

LEMMA 13. *Assume Setting A, and suppose $\bar{x} = 0$. Then, the following must hold:*

$$\|\tilde{s}_i\|_2^2 \leq \left(1 - \frac{\bar{\lambda}_{\min} \delta}{L} \right) \max_{j=1,\ldots,k} \|s_j\|_2^2 \;\forall\; i = 1, \ldots, k,$$

where $\tilde{s}_i$ and $\bar{\lambda}_{\min}$ are as defined in (3.17) *and* (4.1)*, respectively.*

Proof. Consider any fixed i. Then,

$$\begin{aligned}
\|\tilde{s}_i\|_2^2 &= \left\| \frac{1}{2 - \bar{\lambda}_i} \mathcal{P}_{\overline{\mathcal{U}}} \left[s_i - \frac{\bar{\lambda}_i}{L} \nabla f_i(s_i) \right] + \sum_{j \neq i} \frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i} \mathcal{P}_{\overline{\mathcal{U}}} \left[s_j - \frac{1}{L} \nabla f_j(s_j) \right] \right\|_2^2 \\
&\qquad \text{(Substitute (3.17))} \\
&\leq \frac{1}{2 - \bar{\lambda}_i} \left\| \mathcal{P}_{\overline{\mathcal{U}}} \left[s_i - \frac{\bar{\lambda}_i}{L} \nabla f_i(s_i) \right] \right\|_2^2 + \sum_{j \neq i} \frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i} \left\| \mathcal{P}_{\overline{\mathcal{U}}} \left[s_j - \frac{1}{L} \nabla f_j(s_j) \right] \right\|_2^2 \\
&\qquad \text{(Convexity)} \\
&\leq \frac{1}{2 - \bar{\lambda}_i} \left(1 - \frac{\bar{\lambda}_i \delta}{L} \right) \|s_i\|_2^2 + \left(1 - \frac{\delta}{L} \right) \sum_{j \neq i} \frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i} \|s_j\|_2^2 \\
&\qquad \text{(Theorem 12 where } \tilde{L} = \tfrac{L}{\bar{\lambda}_i}\text{)}
\end{aligned}$$

$$\leq \left(1 - \frac{\bar{\lambda}_i \delta}{L}\right) \left[\frac{1}{2 - \bar{\lambda}_i} \|s_i\|_2^2 + \sum_{j \neq i} \frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i} \|s_j\|_2^2\right]$$
$$\leq \left(1 - \frac{\bar{\lambda}_{\min} \delta}{L}\right) \max_{j=1,\ldots,k} \|s_j\|_2^2.$$

This completes the proof. ∎

THEOREM 14 (survey-norm contraction). *Assume Setting A. Then the following must hold:*

$$\max_{j=1,\ldots,k} \left\|s_j^+ - \bar{x}\right\|_2^2 \leq \left(1 - \frac{\bar{\lambda}_{\min} \delta}{L}\right) \max_{j=1,\ldots,k} \|s_j - \bar{x}\|_2^2 + \mathrm{O}\left(\left\|\mathcal{S} - \bar{\mathcal{X}}\right\|^3\right),$$

where $\bar{\lambda}_{\min}$ *is defined as in* (4.1).

Proof. Without loss of generality, assume $\bar{x} = 0$. Consider any fixed i. By Lemma 13, $\tilde{s}_i = \mathrm{O}(\|\mathcal{S}\|)$. Then,

$$\begin{aligned}
\left\|s_i^+\right\|_2^2 &= \left\|\tilde{s}_i + \mathrm{O}\left(\|\mathcal{S}\|^2\right)\right\|_2^2 \text{ (Theorem 11)} \\
&= \|\tilde{s}_i\|_2^2 + 2\left\langle \tilde{s}_i, \mathrm{O}\left(\|\mathcal{S}\|^2\right)\right\rangle + \mathrm{O}\left(\|\mathcal{S}\|^4\right) \\
&\leq \left(1 - \frac{\bar{\lambda}_{\min} \delta}{L}\right) \max_{j=1,\ldots,k} \|s_j\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right) \text{ (Lemma 13).}
\end{aligned}$$

The right-hand side is independent of i, so taking the max over i on the left-hand side leads to the desired result. We can deduce the general $\bar{x} \neq 0$ case with the change of variables $x \leftarrow (x - \bar{x})$. ∎

Theorem 14 implies that for input surveys $\mathcal{S}$, sufficiently close to $\bar{\mathcal{X}}$, survey descent outputs a survey $\mathcal{S}^+$ that is strictly *closer* to $\bar{\mathcal{X}}$. As a result, the feasibility and validity guarantees of Theorems 5 and 7 will continue to hold in repeated applications of survey descent iterations, allowing us to repeat the procedure indefinitely. Letting $\mathbf{N}$ denote the nonnegative natural numbers, we formalize these ideas below.

PROCEDURE B (repeated iterations of survey descent). *For the strong* $\mathcal{C}^2$ *max function objective* f*, assume* $\mathcal{S}^0$ *is a valid initial survey sufficiently close to* $\bar{\mathcal{X}}$*, and initialize* $t = 0$*. Iterate the following steps:*

1. *Solve the survey descent iteration* $\{(P_i^{\mathcal{S}^t})\}_{i=1}^k$ *and denote the output* $\mathcal{S}^{t+1} \equiv (\mathcal{S}^t)^+$.
2. *Increment* $t \leftarrow t + 1$.

COROLLARY 15 (well-definedness of survey descent repetitions). *Procedure B is well defined: for each* $t = 0, 1, 2\ldots$*, the survey descent iteration* $\{(P_i^{\mathcal{S}^t})\}_{i=1}^k$ *is feasible and produces a valid output survey.*

Proof. For $\mathcal{S}^0$ sufficiently close to $\bar{\mathcal{X}}$, Theorem 14 inductively implies that $\|\mathcal{S}^{t+1} - \bar{\mathcal{X}}\| < \|\mathcal{S}^t - \bar{\mathcal{X}}\|$ for all $t \in \mathbf{N}$. In other words, output surveys will always move closer to $\bar{\mathcal{X}}$, and Theorems 5 and 7 continue to hold for survey descent iterations $\{(P_i^{\mathcal{S}^t})\}_{i=1}^k$ for all $t \in \mathbf{N}$. Thus, Theorem 5 implies the feasibility of all survey descent iterations, while Theorem 7 implies the validity of output surveys. ∎

OBSERVATION 16. *The proof of Corollary* 15 *more generally shows that, for all* $t \in \mathbf{N}$*, all surveys* $\mathcal{S}^t$ *will remain sufficiently close to* $\bar{\mathcal{X}}$*. Thus, the conclusions from Theorems* 5–14 *apply to every iteration of Procedure B.*

Having established that the repeated application of survey descent is locally well defined, we now deduce the local Q-linear convergence induced by this procedure.

THEOREM 17 (Q-linear convergence of survey points). *Procedure B satisfies the following property:*

$$\limsup_{t\to\infty} \frac{\left\|\mathcal{S}^{t+1}-\bar{\mathcal{X}}\right\|^2}{\left\|\mathcal{S}^{t}-\bar{\mathcal{X}}\right\|^2} \leq \left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right), \tag{4.2}$$

where $\bar{\lambda}_{\min}$ *is defined as in* (4.1).

Proof. By Theorem 14 and Observation 16, there exists $K>0$ such that

$$\left\|\mathcal{S}^{t+1}-\bar{\mathcal{X}}\right\|^2 \leq \left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)\left\|\mathcal{S}^{t}-\bar{\mathcal{X}}\right\|^2 + K\left\|\mathcal{S}^{t}-\bar{\mathcal{X}}\right\|^3 \ \forall\, t\in\mathbf{N}.$$

Corollary 15 implies that $\mathcal{S}^t$ is a valid survey. By Definition 4, $\bar{\mathcal{X}}$ is not a valid survey because $f_i(\bar{x})=f_j(\bar{x})$ for all i,j. Thus, $\mathcal{S}^t\neq\bar{\mathcal{X}}$, so $\|\mathcal{S}^t-\bar{\mathcal{X}}\|^2$ is nonzero and we can divide both sides by it. We deduce

$$\frac{\left\|\mathcal{S}^{t+1}-\bar{\mathcal{X}}\right\|^2}{\left\|\mathcal{S}^{t}-\bar{\mathcal{X}}\right\|^2} \leq \left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right) + K\left\|\mathcal{S}^{t}-\bar{\mathcal{X}}\right\| \ \forall\, t\in\mathbf{N}.$$

For $\mathcal{S}^0$ sufficiently close to $\bar{\mathcal{X}}$, this inequality guarantees that $\|\mathcal{S}^t-\bar{\mathcal{X}}\|$ monotonically decreases with t and, in particular, $\left\|\mathcal{S}^t-\bar{\mathcal{X}}\right\|<\frac{\bar{\lambda}_{\min}\delta}{LK}$ for all $t\in\mathbf{N}$. Thus, $\frac{\left\|\mathcal{S}^{t+1}-\bar{\mathcal{X}}\right\|}{\left\|\mathcal{S}^{t}-\bar{\mathcal{X}}\right\|}<1$ implying $\lim_{t\to\infty}\|\mathcal{S}^t-\bar{\mathcal{X}}\|=0$. Taking the lim-sup with $t\to\infty$ on both sides of the above inequality then gives the desired result. ∎

Combining Theorem 17 with the L-smoothness assumptions (Definition 3), we can immediately establish a loose R-linear convergence result on the objective values.

COROLLARY 18 (R-linear convergence of objective values, weak version). *Consider Procedure B. Then, for any* $K>\max_j\|\nabla f_j(\bar{x})\|_2$,

$$f\left(s_i^t\right)-f(\bar{x}) \leq K\left\|\mathcal{S}^t-\bar{\mathcal{X}}\right\| \ \forall\, i=1,\ldots,k, \text{ and } t\in\mathbf{N}.$$

Consequently, survey objective values $\{f(s_i^t)\}_{t\in\mathbf{N}}$ *converge R-linearly to* $f(\bar{x})$ *for all* i.

Proof. The proof follows from the L-smoothness of the f_i-components and a routine application of the Cauchy–Schwarz inequality. R-linear convergence follows from the Q-linear convergence of $\{\|\mathcal{S}^t-\bar{\mathcal{X}}\|\}_{t\in\mathbf{N}}$ (Theorem 17). ∎

Compared to classical GD convergence guarantees on smooth objectives [2, Theorem 10.29], Corollary 18 is rather weak: Firstly, the upper-bounding sequence is $\{\|S^t-\bar{\mathcal{X}}\|\}_{t\in\mathbf{N}}$ rather than the tighter $\{\|S^t-\bar{\mathcal{X}}\|^2\}_{t\in\mathbf{N}}$ seen in the smooth GD case; secondly, the constant K depends on $\max_j\|\nabla f_j(\bar{x})\|_2$, which does not appear in the smooth GD case. A stronger function-value convergence guarantee for survey descent—more structurally similar to the aforementioned GD results—is indeed possible.

THEOREM 19 (R-linear vonvergence of objective values, strong version). *Consider Procedure B. Fix* $\bar{K}>\frac{5L}{\bar{\lambda}_{\min}}\left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)$. *Then,*

$$f\left(s_i^{t+1}\right)-f(\bar{x}) \leq \bar{K}\left\|\mathcal{S}^t-\bar{\mathcal{X}}\right\|^2 \ \forall\, i=1,\ldots,k, \text{ and } t\in\mathbf{N}.$$

Consequently, survey objective values $\{f(s_i^t)\}_{t\in\mathbf{N}}$ *converge R-linearly to* $f(\bar{x})$ *for all* i.

Proof. The proof follows from combining Theorem 11, Theorem 12, and the L-smoothness and δ-strong convexity assumptions on the components $\{f_i\}_{i=1}^k$. We defer a detailed description to Appendix A. ∎

Note that the rates of linear convergence presented in Theorems 17 and 19 are conservative. Experiments suggest that these rate bounds are far from tight.

5. Implementing survey descent. Stepping back from our theoretical analysis on strong $\mathcal{C}^2$ max functions f, we conclude with two remarks on the potential implementation of survey descent on general nonsmooth objectives h (Definition 1).

Remark 20 (informal heuristic for survey initialization). We sketch an informal idea. First, run some initializing iterations of a standard method such as BFGS or a subgradient or proximal bundle method, and collect the gradients at these iterates. (As discussed, these iterates are typically points of differentiability.) In practice, the iterates explore the nonsmooth landscape effectively. In particular, on max functions, they "bounce" between the active functions near the minimizer: for BFGS, see our Figure 5 and [31, Figure 5.5,6], and for the subgradient method, see the figures and discussion on "oscillations" in [4]. As discussed, the convex hulls of nearby iterate gradients thus approximate the subdifferential at optimality, so we can estimate its dimension d via singular value decomposition. An initialization heuristic could then select a survey of $k = d + 1$ iterates with robustly affinely independent gradients. Implementing such heuristics is intricate: we defer further discussion to future work.

Remark 21 (possible acceleration of survey kescent). First, note that we can solve the k constituent subproblems of survey descent (Definition 1) in parallel. Second, when all subproblems are feasible, we found in exploratory experiments that the following adjustment expedites survey descent's empirical convergence:

> *For all i, only update $s_i \leftarrow s_i^+$ when $h(s_i^+) < h(s_i)$; otherwise, keep s_i the same in the subsequent survey.*

This also tends to make survey descent compatible with a wider variety of initializing heuristics. For simplicity, we omit this additional enhancement from Definition 1 and do not apply it in this paper's experiments.

Appendix A. Stronger R-linear convergence of function values (Theorem 19). In section 4, we analyzed the distance of the survey descent updates, $\mathcal{S}^+ = \{s_i^+\}_{i=1}^k$, from the objective minimizer by examining the set of reference points $\{\tilde{s}_i\}_{i=1}^k$—which coincides with $\mathcal{S}^+$ up to $\mathrm{O}(\|\mathcal{S}\|^2)$ terms and is defined in (3.17). This appendix will prove Theorem 19—showing R-linear function-value convergence—through an analogous strategy. To simplify the derivations, we adopt the following notation:

$$\mathring{s}_i = \mathcal{P}_{\overline{\mathcal{U}}}\left[s_i - \frac{1}{L}\nabla f_i(s_i)\right] \text{ and } \hat{s}_i = \mathcal{P}_{\overline{\mathcal{U}}}\left[s_i - \frac{\bar{\lambda}_i}{L}\nabla f_i(s_i)\right] \ \forall\ i = 1, \dots, k. \tag{A.1}$$

In this notation, (3.17) simplifies to

$$\tilde{s}_i \equiv \frac{1}{2 - \bar{\lambda}_i}\hat{s}_i + \sum_{j \neq i} \frac{\bar{\lambda}_j}{2 - \bar{\lambda}_i}\mathring{s}_j. \tag{A.2}$$

For any fixed i, we will prove Theorem 19 in three steps:

1. Lemma **22**: Bounding the deviation between $f_i(\mathring{s}_i)$ and $f(\mathring{s}_i)$ because $\mathring{s}_i$ does *not* necessarily satisfy $f_i(\mathring{s}_i) = f(\mathring{s}_i)$. Similarly, we also bound the difference between $f_i(\hat{s}_i)$ and $f(\hat{s}_i)$.

2. Lemma **23**: Using convexity and Theorem 12 to bound $f(\tilde{s}_i)$.
3. Lemma **24**: Bounding the difference between $f(s_i^+)$ and $f(\tilde{s}_i)$.

We will again need to invoke the Taylor expansion (3.5) to construct our bounds. However, we require a stronger version of Taylor's theorem than that in Lemma 8—using "Little-Oh" rather than "Big-Oh." In particular, for a mapping $g : \mathbf{E}\to\mathbf{F}$ between two Euclidean spaces and letting $\|\cdot\|^p$ denote an arbitrary Euclidean norm raised to the pth power, we use the notation $g(u) = \mathrm{o}(\|u\|^p)$ to indicate the property of g that—given *any* constant $K > 0$—there exists $U_K > 0$ such that $|g(u)| \le K\|u\|^p$ for all $\|u\| \le U_K$. By itself, we let $\mathrm{o}(\|u\|^p)$ denote an element of the class of all functions with this property. Then, an identical argument as that in Lemma 8 shows that the residual function in (3.5) satisfies

$$r_i(x) = \mathrm{o}\left(\|x\|^2\right) \text{ and } \nabla r_i(x) = \mathrm{o}\left(\|x\|\right). \tag{A.3}$$

We now prove our main result.

LEMMA 22. *Assume Setting A, and suppose $\bar{x} = 0$. Then, the following inequalities hold:*

$$|f_i(\mathring{s}_i) - f(\mathring{s}_i)| \le \frac{L}{2}\left(1 - \frac{\delta}{L}\right)\|s_i\|_2^2 + \mathrm{o}\left(\|s_i\|_2^2\right),$$
$$|f_i(\hat{s}_i) - f(\hat{s}_i)| \le \frac{L}{2\bar{\lambda}_i}\left(1 - \frac{\bar{\lambda}_i\delta}{L}\right)\|s_i\|_2^2 + \mathrm{o}\left(\|s_i\|_2^2\right).$$

Proof. Using the Taylor expansion (3.5), observe for all pairs i, j that

$$\begin{aligned}
|f_i(\mathring{s}_i) - f_j(\mathring{s}_i)| &= \left|(a_i - a_j)^T\mathring{s}_i + \frac{1}{2}\mathring{s}_i^T(A_i - A_j)\mathring{s}_i + r_i(\mathring{s}_i) - r_j(\mathring{s}_i)\right| \\
&= \left|\frac{1}{2}\mathring{s}_i^T(A_i - A_j)\mathring{s}_i + r_i(\mathring{s}_i) - r_j(\mathring{s}_i)\right| \text{ (by } \mathring{s}_i \in \overline{\mathcal{U}} \text{ and (3.1))} \\
&\le \frac{L}{2}\|\mathring{s}_i\|_2^2 + \mathrm{o}\left(\|\mathring{s}_i\|_2^2\right) \text{ (by } L\text{-smoothness and (A.3))} \\
&\le \frac{L}{2}\left(1 - \frac{\delta}{L}\right)\|s_i\|_2^2 + \mathrm{o}\left(\|s_i\|_2^2\right) \text{ (Theorem 12).}
\end{aligned}$$

Observe the right-hand side is independent of j, so the above still holds after replacing f_j with $f = \max_j f_j$, which gives our desired result. An analogous argument for $\hat{s}_i$, which replaces $\frac{1}{L}$ with $\frac{\bar{\lambda}_i}{L}$, shows

$$|f_i(\hat{s}_i) - f(\hat{s}_i)| \le \frac{L}{2\bar{\lambda}_i}\left(1 - \frac{\bar{\lambda}_i\delta}{L}\right)\|s_i\|_2^2 + \mathrm{o}\left(\|s_i\|_2^2\right).$$

This completes the proof. ∎

LEMMA 23. *Assume Setting A, and suppose $\bar{x} = 0$ and $f(\bar{x}) = 0$. Then, the following inequality holds:*

$$f(\tilde{s}_i) \le \frac{L}{\bar{\lambda}_{\min}}\left(1 - \frac{\delta\bar{\lambda}_{\min}}{L}\right)\max_{j=1,\ldots,k}\|s_j\|_2^2 + \mathrm{o}\left(\|\mathcal{S}\|^2\right), \tag{A.4}$$

where $\tilde{s}_i$ and $\bar{\lambda}_{\min}$ are as defined in (3.17) *and* (4.1)*, respectively.*

Proof. Using (A.2), Theorem 11, and convexity, we deduce

$$f(\tilde{s}_i) = f\left(\frac{1}{2-\bar{\lambda}_i}\hat{s}_i + \sum_{j\neq i}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_i}\mathring{s}_j\right) \leq \frac{1}{2-\bar{\lambda}_i}f(\hat{s}_i) + \sum_{j\neq i}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_i}f(\mathring{s}_j). \tag{A.5}$$

Define the quantity

$$\Delta \equiv \frac{L}{2\bar{\lambda}_{\min}}\left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)\left(\frac{1}{2-\bar{\lambda}_i}\|s_i\|_2^2 + \sum_{j\neq i}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_i}\|s_j\|_2^2\right).$$

Applying Lemma 22 to (A.5) combined with the facts that $\bar{\lambda}_{\min} \leq \bar{\lambda}_j \leq 1$ for all j gives

$$f(\tilde{s}_i) \leq \underbrace{\frac{1}{2-\bar{\lambda}_i}f_i(\hat{s}_i) + \sum_{j\neq i}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_i}f_j(\mathring{s}_j)}_{(*)} + \Delta + \mathrm{o}\left(\|\mathcal{S}\|^2\right). \tag{A.6}$$

Theorem 12 implies the following bound on $(*)$:

$$(*) \leq \frac{1}{2-\bar{\lambda}_i}\frac{L}{2\bar{\lambda}_i}\left(1-\frac{\delta\bar{\lambda}_i}{L}\right)\|s_i\|^2 + \frac{L}{2}\left(1-\frac{\delta}{L}\right)\sum_{j\neq i}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_i}\|s_j\|_2^2 \leq \Delta.$$

Substituting the above into (A.6) gives the desired

$$\begin{aligned} f(\tilde{s}_i) &\leq 2\Delta + \mathrm{o}\left(\|\mathcal{S}\|_2^2\right) \\ &= \frac{L}{\bar{\lambda}_{\min}}\left(1-\frac{\delta\bar{\lambda}_{\min}}{L}\right)\left[\frac{1}{2-\bar{\lambda}_i}\|s_i\|^2 + \sum_{j\neq i}\frac{\bar{\lambda}_j}{2-\bar{\lambda}_i}\|s_j\|_2^2\right] + \mathrm{o}\left(\|\mathcal{S}\|^2\right) \\ &\leq \frac{L}{\bar{\lambda}_{\min}}\left(1-\frac{\delta\bar{\lambda}_{\min}}{L}\right)\max_{j=1,\ldots,k}\|s_j\|_2^2 + \mathrm{o}\left(\|\mathcal{S}\|^2\right), \end{aligned}$$

which is our desired result. ∎

LEMMA 24. *Assume Setting A, and suppose* $\bar{x} = 0$ *and* $f(\bar{x}) = 0$. *Then, the following inequality holds:*

$$f\left(s_i^+\right) - f(\tilde{s}_i) \leq \frac{4L}{\bar{\lambda}_{\min}}\left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)\max_{j=1,\ldots,k}\|s_i\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right),$$

where $\tilde{s}_i$ *and* $\bar{\lambda}_{\min}$ *are as defined in* (3.17) *and* (4.1), *respectively.*

Proof. Consider any fixed i. By Theorem 5, all constraints of $(P_i^{\mathcal{S}})$ are tight, so s_i^+ satisfies the constraints of Observation 6 with equality for all $j \neq i$:

$$\ell_{s_j}^{f_j}\left(s_i^+\right) + \frac{L}{2}\left\|s_i^+ - s_j\right\|_2^2 = \ell_{s_i}^{f_i}\left(s_i^+\right). \tag{A.7}$$

Define $\xi_i \equiv s_i^+ - \tilde{s}_i$. In other words, ξ_i is the $\mathrm{O}(\|\mathcal{S}\|^2)$ term in (3.16). Consider any fixed j. Substituting $s_i^+ = \tilde{s}_i + \xi_i$ into (A.7) and expanding the quadratic term gives

$$\ell_{s_j}^{f_j}(\tilde{s}_i) + \nabla f_j(s_j)^T\xi_i + \frac{L}{2}\|\tilde{s}_i - s_j\|_2^2 + L\langle\tilde{s}_i - s_j, \xi_i\rangle + \frac{L}{2}\|\xi_i\|^2 = \ell_{s_i}^{f_i}(\tilde{s}_i) + \nabla f_i(s_i)^T\xi_i.$$

Applying the Taylor expansion (3.5) to $\nabla f_i(s_i)$ and $\nabla f_j(s_j)$ gives

$$\begin{aligned}&\ell_{s_j}^{f_j}(\tilde{s}_i)+\left(a_j+A_js_j+\nabla r_j(s_j)\right)^T\xi_i+\frac{L}{2}\left\|\tilde{s}_i-s_j\right\|_2^2+L\left\langle\tilde{s}_i-s_j,\xi_i\right\rangle+\frac{L}{2}\left\|\xi_i\right\|^2\\&\quad=\ell_{s_i}^{f_i}(\tilde{s}_i)+\left(a_i+A_is_i+\nabla r_i(s_i)\right)^T\xi_i.\end{aligned}$$

Note that the terms $\nabla r_j(s_j)$, $\nabla r_i(s_i)$, and $\tilde{s}_i$ are all $\mathrm{O}(\|\mathcal{S}\|)$ by Lemma 8 and (3.17), and so is s_i by definition. Thus, collecting $\mathrm{O}(\|\mathcal{S}\|^3)$ terms simplifies the above to

$$\ell_{s_j}^{f_j}(\tilde{s}_i)+a_j^T\xi_i+\frac{L}{2}\left\|\tilde{s}_i-s_j\right\|_2^2+\mathrm{O}\left(\|\mathcal{S}\|^3\right)=\ell_{s_i}^{f_i}(\tilde{s}_i)+a_i^T\xi_i. \tag{A.8}$$

Multiplying (A.8) by $\bar{\lambda}_j$, summing over $j\neq i$, and using identities (2.5) and (2.6) gives

$$\begin{aligned}&\sum_{j\neq i}\bar{\lambda}_j\ell_{s_j}^{f_j}(\tilde{s}_i)-\bar{\lambda}_ia_i^T\xi_i+\frac{L}{2}\sum_{j\neq i}\left(\bar{\lambda}_j\left\|\tilde{s}_i-s_j\right\|_2^2\right)+\mathrm{O}\left(\|\mathcal{S}\|^3\right)\\&\quad=\left(1-\bar{\lambda}_i\right)\ell_{s_i}^{f_i}(\tilde{s}_i)+\left(1-\bar{\lambda}_i\right)a_i^T\xi_i.\end{aligned}$$

Adding $\bar{\lambda}_i\ell_{s_i}^i(\tilde{s}_i)$ to both sides and again applying Lemma 8 leads to

$$\begin{aligned}&\sum_{j=1}^k\bar{\lambda}_j\ell_{s_j}^{f_j}(\tilde{s}_i)+\frac{L}{2}\sum_{j\neq i}\left(\bar{\lambda}_j\left\|\tilde{s}_i-s_j\right\|_2^2\right)+\mathrm{O}\left(\|\mathcal{S}\|^3\right)\\&\quad=\ell_{s_i}^{f_i}(\tilde{s}_i)+a_i^T\xi_i=\ell_{s_i}^{f_i}\left(s_i^+\right)\underbrace{-\left(A_is_i+\nabla r_i(s_i)\right)^T\xi_i}_{\mathrm{O}(\|\mathcal{S}\|^3)}.\end{aligned}$$

Again collecting $\mathrm{O}(\|\mathcal{S}\|^3)$ terms and adding a nonnegative $\frac{\bar{\lambda}_iL}{2}\|\tilde{s}_i-s_i\|_2^2$ term to the left-hand side leads to the following inequality:

$$\underbrace{\sum_{j=1}^k\bar{\lambda}_j\ell_{s_j}^{f_j}(\tilde{s}_i)}_{(\maltese)}+\frac{L}{2}\sum_{j=1}^k\left(\bar{\lambda}_j\left\|\tilde{s}_i-s_j\right\|_2^2\right)+\mathrm{O}\left(\|\mathcal{S}\|^3\right)\geq\ell_{s_i}^{f_i}\left(s_i^+\right) \tag{A.9}$$

For ($\maltese$), the assumed δ-strong convexity on f_j (Definition 3) implies

$$\sum_{j=1}^k\bar{\lambda}_j\ell_{s_j}^{f_j}(\tilde{s}_i)\leq\sum_{j=1}^k\bar{\lambda}_j\left(f_j(\tilde{s}_i)-\frac{\delta}{2}\left\|\tilde{s}_i-s_j\right\|_2^2\right)\leq\underbrace{\max_{j=1,\ldots,k}f_j(\tilde{s}_i)}_{=f(\tilde{s}_i)}-\frac{\delta}{2}\sum_{j=1}^k\bar{\lambda}_j\left\|\tilde{s}_i-s_j\right\|_2^2, \tag{A.10}$$

where the rightmost inequality follows from (2.5). Substituting (A.10) into (A.9) leads to

$$\begin{aligned}&f(\tilde{s}_i)+\frac{L}{2}\left(1-\frac{\delta}{L}\right)\sum_{j=1}^k\left(\bar{\lambda}_j\left\|\tilde{s}_i-s_j\right\|_2^2\right)+\mathrm{O}\left(\|\mathcal{S}\|^3\right)\\&\quad\geq\ell_{s_i}^{f_i}\left(s_i^+\right)\geq f_i\left(s_i^+\right)-\frac{L}{2}\left\|s_i^+-s_i\right\|_2^2,\end{aligned} \tag{A.11}$$

where the second inequality follows from the L-smoothness of f_i (Definition 3). By Theorem 7, s_i^+ is a valid survey point, so $f_i(s_i^+) = f(s_i^+)$. Substituting this into (A.11) and rearranging gives

$$f\left(s_i^+\right) \leq f\left(\tilde{s}_i\right) + \frac{L}{2}\left(1-\frac{\delta}{L}\right)\underbrace{\sum_{j=1}^{k}\left(\bar{\lambda}_j \left\|\tilde{s}_i - s_j\right\|_2^2\right)}_{(\dagger)} + \frac{L}{2}\underbrace{\left\|s_i^+ - s_i\right\|_2^2}_{(\star)} + \mathrm{O}\left(\|\mathcal{S}\|^3\right). \tag{A.12}$$

We use Jensen's inequality to bound $(\star)$:

$$\begin{aligned}(\star) &\leq \max_{j=1,\ldots,k}\left\|s_i^+ - s_j\right\|_2^2 \\ &= \max_{j=1,\ldots,k}\left\|\frac{1}{2}\left(2s_i^+\right) + \frac{1}{2}\left(-2s_j\right)\right\|_2^2 \leq \frac{1}{2}\left\|\left(2s_i^+\right)\right\|_2^2 + \frac{1}{2}\max_{j=1,\ldots,k}\|(-2s_j)\|_2^2.\end{aligned}$$

Hence,

$$\begin{aligned}(\star) &\leq 2\left(\left\|s_i^+\right\|_2^2 + \max_{j=1,\ldots,k}\|s_j\|_2^2\right) \\ &= 2\left(2-\frac{\bar{\lambda}_{\min}\delta}{L}\right)\max_{j=1,\ldots,k}\|s_j\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right) \text{ (Theorem 14)} \\ &\leq 4\max_{j=1,\ldots,k}\|s_j\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right).\end{aligned}$$

An analogous argument (using Lemma 13 on $\tilde{s}_i$ instead of Theorem 14 on s_i^+) gives

$$(\dagger) \leq \max_{j=1,\ldots,k}\left\|\tilde{s}_i - s_j\right\|_2^2 \leq 4\max_{j=1,\ldots,k}\|s_j\|_2^2.$$

Substituting these bounds on $(\dagger)$ and $(\star)$ into (A.12), we deduce

$$\begin{aligned}f\left(s_i^+\right) &\leq f\left(\tilde{s}_i\right) + 2L\left(1-\frac{\delta}{L}\right)\max_{j=1,\ldots,k}\|s_i\|_2^2 + 2L\max_{j=1,\ldots,k}\|s_i\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right) \\ &= f\left(\tilde{s}_i\right) + 4L\left(1-\frac{\delta}{2L}\right)\max_{j=1,\ldots,k}\|s_i\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right).\end{aligned}$$

If $k=1$, survey descent reduces to GD, and the result is trivial from [2, Theorem 10.29] or [38, Theorem 2.1.15]. When $k \geq 2$, the property (2.5) implies $\bar{\lambda}_{\min} \leq \frac{1}{k} \leq \frac{1}{2}$. Thus, $\left(1-\frac{\delta}{2L}\right) \leq \left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)$. We then deduce

$$f\left(s_i^+\right) \leq f\left(\tilde{s}_i\right) + \frac{4L}{\bar{\lambda}_{\min}}\left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)\max_{j=1,\ldots,k}\|s_i\|_2^2 + \mathrm{O}\left(\|\mathcal{S}\|^3\right).$$

This completes the proof. ∎

THEOREM 19 (R-linear convergence of objective values, strong version). *Consider Procedure B. Fix* $\bar{K} > \frac{5L}{\bar{\lambda}_{\min}}\left(1-\frac{\bar{\lambda}_{\min}\delta}{L}\right)$. *Then,*

$$f\left(s_i^{t+1}\right) - f(\bar{x}) \leq \bar{K}\left\|\mathcal{S}^t - \bar{\mathcal{X}}\right\|^2 \ \forall\ i=1,\ldots,k, \text{ and } t \in \mathbf{N}.$$

Consequently, survey objective values $\{f(s_i^t)\}_{t\in\mathbf{N}}$ *converge R-linearly to* $f(\bar{x})$ *for all* i.

Proof. Without loss of generality, assume $\bar{x} = 0$ and $f(\bar{x}) = 0$. Consider any fixed i and t. For simplicity, adopt the notation $s_i = s_i^t$ and $s_i^+ = s_i^{t+1}$. Adding the inequalities in Lemmas 23 and 24 gives

$$f\left(s_i^+\right) \leq \frac{5L}{\bar{\lambda}_{\min}}\left(1 - \frac{\bar{\lambda}_{\min}\delta}{L}\right) \max_{j=1,\ldots,k} \|s_j\|_2^2 + \underbrace{\mathrm{o}\left(\|\mathcal{S}\|^2\right) + \mathrm{O}\left(\|\mathcal{S}\|^3\right)}_{=\mathrm{o}(\|\mathcal{S}\|^2)}.$$

Then, for any $K > 0$ and for all $\mathcal{S}$ sufficiently close to $\bar{\mathcal{X}}$, we have

$$f\left(s_i^+\right) \leq \frac{5L}{\bar{\lambda}_{\min}}\left(1 - \frac{\bar{\lambda}_{\min}\delta}{L}\right)\|\mathcal{S}\|^2 + K\|\mathcal{S}\|^2 = \left(\frac{5L}{\bar{\lambda}_{\min}}\left(1 - \frac{\bar{\lambda}_{\min}\delta}{L}\right) + K\right)\|\mathcal{S}\|^2.$$

Denote $\bar{K} \equiv \frac{5L}{\bar{\lambda}_{\min}}\left(1 - \frac{\bar{\lambda}_{\min}\delta}{L}\right) + K > 0$. Since K can be any strictly positive constant, choosing K is equivalent to choosing an arbitrary $\bar{K} > \frac{5L}{\bar{\lambda}_{\min}}\left(1 - \frac{\bar{\lambda}_{\min}\delta}{L}\right)$. Under such a choice,

$$f\left(s_i^+\right) \leq \bar{K}\|\mathcal{S}\|^2,$$

which is our desired result. The general case follows from applying the change of variables $x \leftarrow (x - \bar{x})$ and function-value shift $f(\cdot) \leftarrow (f(\cdot) - f(\bar{x}))$. ∎

REFERENCES

[1] F. Atenas, C. Sagastizábal, P. J. Silva, and M. Solodov, *A Unified Analysis of Descent Sequences in Weakly Convex Optimization, Including Convergence Rates for Bundle Methods*, Technical report, Optimization Online, 2021, https://optimization-online.org/?p=18426.

[2] A. Beck, *First-Order Methods in Optimization*, SIAM, Philadelphia, 2017.

[3] P. Bianchi, W. Hachem, and S. Schechtman, *Convergence of constant step stochastic gradient descent for non-smooth non-convex functions*, Set-Valued Var. Anal., 30 (2022), pp. 1117–1147.

[4] J. Bolte, E. Pauwels, and R. Rios-Zertuche, *Long Term Dynamics of the Subgradient Method for Lipschitz Path Differentiable Functions*, arXiv preprint, arXiv:2006.00098, 2020.

[5] A. Borghetti, A. Frangioni, F. Lacalandra, and C. A. Nucci, *Lagrangian heuristics based on disaggregated bundle methods for hydrothermal unit commitment*, IEEE Trans. Power Syst., 18 (2003), pp. 313–323.

[6] J. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*, Springer Science & Business Media, New York, 2010.

[7] J. V. Burke, F. E. Curtis, A. S. Lewis, M. L. Overton, and L. E. Simões, *Gradient sampling methods for nonsmooth optimization*, in Numerical Nonsmooth Optimization: State of the Art Algorithms, A. M. Bagirov, M. Gaudioso, N. Karmitsa, M. M. Mäkelä, and S. Taheri, eds., Springer, Cham, 2020, pp. 201–225.

[8] Y. Carmon, A. Jambulapati, Y. Jin, and A. Sidford, *Thinking inside the ball: Near-optimal minimization of the maximal loss*, in Proceedings of the Conference on Learning Theory, 2021, pp. 866–882.

[9] F. H.Clarke, *Optimization and Nonsmooth Analysis*, SIAM, Philadelphia, 1990.

[10] A. Daniilidis and P. Georgiev, *Approximate convexity and submonotonicity*, J. Math. Anal. Appl., 291 (2004), pp. 292–301.

[11] D. Davis, D. Drusvyatskiy, and L. Jiang, *Subgradient Methods Near Active Manifolds: Saddle Point Avoidance, Local Convergence, and Asymptotic Normality*, arXiv preprint, arXiv:2108.11832, 2021.

[12] D. Davis and L. Jiang, *A Nearly Linearly Convergent First-Order Method for Nonsmooth Functions with Quadratic Growth*, arXiv preprint, arXiv:2205.00064, 2022.

[13] W. de Oliveira and C. Sagastizábal, *Bundle methods in the XXIst century: A bird's-eye view*, Pesqui. Oper., 34 (2014), pp. 647–670.

[14] W. de Oliveira and C. Sagastizábal, *Level bundle methods for oracles with on-demand accuracy*, Optim. Methods Softw., 29 (2014), pp. 1180–1209.
[15] M. Diaz and B. Grimmer, *Optimal Convergence Rates for the Proximal Bundle Method*, arXiv preprint, arXiv:2105.07874, 2021.
[16] Y. Du and A. Ruszczynski, *Rate of convergence of the bundle method*, J. Optim. Theory Appl., 173 (2017), pp. 908–922.
[17] J. C. Duchi and F. Ruan, *Asymptotic optimality in stochastic optimization*, Ann. Statist., 49 (2021), pp. 21–48.
[18] G. Emiel and C. Sagastizábal, *Incremental-like bundle methods with application to energy planning*, Comput. Optim. Appl., 46 (2010), pp. 305–332.
[19] R. Garmanjani, D. Júdice, and L. N. Vicente, *Trust-region methods without using derivatives: Worst case complexity and the nonsmooth case*, SIAM J. Optim., 26 (2016), pp. 1987–2011.
[20] X. Y. Han and A. S. Lewis, *Survey Descent: A Multipoint Generalization of Gradient Descent for Nonsmooth Optimization*, arXiv preprint, arXiv:2111.15645, 2021.
[21] W. Hare and J. Nutini, *A derivative-free approximate gradient sampling algorithm for finite minimax problems*, Comput. Optim. Appl., 56 (2013), pp. 1–38.
[22] A. Juditsky and A. Nemirovski, *First order methods for nonsmooth convex large-scale optimization,* I*: General purpose methods*, Optim. Mach. Learn., 30 (2011), pp. 121–148.
[23] K. A. Khan, J. Larson, and S. M. Wild, *Manifold sampling for optimization of nonconvex functions that are piecewise linear compositions of smooth components*, SIAM J. Optim., 28 (2018), pp. 3001–3024.
[24] K. Kiwiel, *A phase* I*—phase* II *method for inequality constrained minimax problems*, Control Cybern., 12 (1983), pp. 55–75.
[25] K. C. Kiwiel, *Proximal level bundle methods for convex nondifferentiable optimization, saddle-point problems and variational inequalities*, Math. Program., 69 (1995), pp. 89–109.
[26] K. C. Kiwiel, *Efficiency of proximal bundle methods*, J. Optim. Theory Appl., 104 (2000), pp. 589–603.
[27] C. Lemarechal, *An extension of Davidon methods to non differentiable problems*, in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Springer, Cham, 1975, pp. 95–109.
[28] C. Lemaréchal, A. Nemirovskii, and Y. Nesterov, *New variants of bundle methods*, Math. Program., 69 (1995), pp. 111–147.
[29] C. Lemaréchal, F. Oustry, and C. Sagastizábal, *The $\mathcal{U}$-Lagrangian of a convex function*, Trans. Amer. Math. Soc., 352 (2000), pp. 711–729.
[30] A. S. Lewis, *Active sets, nonsmoothness, and sensitivity*, SIAM J. Optim., 13 (2002), pp. 702–725.
[31] A. S. Lewis and M. L. Overton, *Nonsmooth Optimization via BFGS*, Technical report, Optimization Online, 2008, https://optimization-online.org/?p=10625.
[32] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Math. Program., 141 (2013), pp. 135–163.
[33] J. Liang and R. D. Monteiro, *A proximal bundle variant with optimal iteration-complexity for a large range of prox stepsizes*, SIAM J. Optim., 31 (2021), pp. 2955–2986.
[34] J. Liang and R. D. Monteiro, *A Unified Analysis of a Class of Proximal Bundle Methods for Solving Hybrid Convex Composite Optimization Problems*, arXiv preprint, arXiv:2110.01084, 2021.
[35] R. Mifflin and C. Sagastizábal, *A $\mathcal{VU}$-algorithm for convex minimization*, Math. Program., 104 (2005), pp. 583–608.
[36] R. Mifflin and C. Sagastizábal, *A science fiction story in nonsmooth optimization originating at IIASA*, Doc. Math., (2012), pp. 291–300.
[37] A. S. Nemirovski and D. B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.
[38] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*, Appl. Optim. 87, Springer Science & Business Media, New York, 2003.
[39] J. Nocedal and S. Wright, *Numerical Optimization*, Springer Science & Business Media, New York, 2006.
[40] D. Noll and P. Apkarian, *Spectral bundle methods for non-convex maximum eigenvalue functions: Second-order methods*, Math. Program., 104 (2005), pp. 729–747.
[41] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, SIAM, Philadelphia, 2000.
[42] F. Oustry, *A second-order bundle method to minimize the maximum eigenvalue function*, Math. Program., 89 (2000), pp. 1–33.

[43] S. M. Robinson, *Linear convergence of epsilon-subgradient descent methods for a class of convex functions*, Math. Program., 86 (1999), pp. 41–50.
[44] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*, Grundlehren Math. Wiss. 317, Springer Science & Business Media, New York, 2009.
[45] C. Sagastizábal, *Divide to conquer: Decomposition methods for energy optimization*, Math. Program., 134 (2012), pp. 187–222.
[46] J. E. Spingarn, *Submonotone subdifferentials of Lipschitz functions*, Trans. Amer. Math. Soc., 264 (1981), pp. 77–89.
[47] C. H. Teo, S. Vishwanathan, A. Smola, and Q. V. Le, *Bundle methods for regularized risk minimization*, J. Mach. Learn. Res., 11 (2010), pp. 311–365.
[48] W. van Ackooij, R. Henrion, A. Möller, and R. Zorgati, *Joint chance constrained programming for hydro reservoir management*, Optim. Eng., 15 (2014), pp. 509–531.
[49] P. Wolfe, *A method of conjugate subgradients for minimizing nondifferentiable functions*, in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Springer, Cham, 1975, pp. 145–173.