Journal of Machine Learning Research 20 (2019) 1-52 Submitted 11/18; Published 7/19

Convergence guarantees for a class of non-convex and non-smooth optimization problems

Koulik Khamaru[⋆] koulik@berkeley.edu
Martin J. Wainwright[⋆,†,‡] wainwrig@eecs.berkeley.edu
Department of Statistics[⋆]
Department of Electrical Engineering & Computer Sciences[†]
University of California Berkeley
Berkeley, CA 94720-1776, USA
Voleon Group[‡]
Berkeley, CA

Editor: Alexander Rakhlin

Abstract

We consider the problem of finding critical points of functions that are non-convex and non-smooth. Studying a fairly broad class of such problems, we analyze the behavior of three gradient-based methods (gradient descent, proximal update, and Frank-Wolfe update). For each of these methods, we establish rates of convergence for general problems, and also prove faster rates for continuous sub-analytic functions. We also show that our algorithms can escape strict saddle points for a class of non-smooth functions, thereby generalizing known results for smooth functions. Our analysis leads to a simplification of the popular CCCP algorithm, used for optimizing functions that can be written as a difference of two convex functions. Our simplified algorithm retains all the convergence properties of CCCP, along with a significantly lower cost per iteration. We illustrate our methods and theory via applications to the problems of best subset selection, robust estimation, mixture density estimation, and shape-from-shading reconstruction.

1. Introduction

Non-convex optimization problems arise frequently in statistical machine learning; examples include the use of non-convex penalties for enforcing sparsity (Fan and Li, 2001; Loh and Wainwright, 2013; Wainwright, 2019), non-convexity in likelihoods in mixture modeling (Yan et al., 2017), and non-convexity in neural network training (Li and Yuan, 2017). Of course, minimizing a non-convex problem is NP-hard in general, but problems that arise in machine learning applications are not constructed in an adversarial manner. Moreover, there have been a number of recent papers demonstrating that all first (and/or second) order critical points have desirable properties for certain statistical problems (e.g. Loh and Wainwright 2013; Ge et al. 2017). Given results of this type, it is often sufficient to find critical points that are first-order (and possibly second-order) stationary. Accordingly, recent years have witnessed an explosion of research on different algorithms for non-convex problems, with the goal of trying to characterize the nature of their fixed points, and their convergence properties.

There is a lengthy literature on non-convex optimization, dating back more than six decades, and rapidly evolving in the present (e.g., see Tuy 1995; Hartman 1959; Horst et al. 2000; Lanckriet and Sriperumbudur 2009; Yuille and Rangarajan 2003; Lee et al. 2016; Bolte et al. 2014; Panageas and Piliouras 2016; Lipp and Boyd 2016; Cartis et al. 2010; Attouch et al. 2010; Gotoh et al. 2017). Perhaps the most straightforward approach to obtaining a first-order critical point is via gradient descent. Under suitable regularity conditions and step size choices, it can be shown that gradient descent can be used to compute first-order critical points. Moreover, with a random initialization and additional regularity conditions, gradient descent converges almost surely to a second-order stationary point (e.g., Lee et al. 2016; Panageas and Piliouras 2016). These results, like much of the currently available theory for (sub)-gradient methods for non-convex problems, involve smoothness conditions on the underlying objectives. In practice, many machine learning problems have non-smooth components; examples include the hinge loss in support vector machines, the rectified linear unit in neural networks, and various types of matrix regularizers in collaborative filtering and recommender systems. Accordingly, a natural goal is to develop subgradient-based techniques that apply to a broader class of non-convex functions, allowing for non-smoothness.

The main contribution of this paper is to provide precisely such a set of techniques, along with non-asymptotic guarantees on their convergence rates. In particular, we study algorithms that can be used to obtain first-order (and in some cases, also second-order) optimal solutions to a relatively broad class of non-convex functions, allowing for non-smoothness in certain portions of the problem. For each sequence $\{x^k\}_{k\geq 0}$ generated by one of our algorithms, we provide non-asymptotic bounds on the convergence rate of the gradient sequence $\{\|\nabla f(x^k)\|_2\}_{k\geq 0}$. Moreover, for functions that satisfy a form of the Kurdaya-Łojasiewicz inequality, we show that our methods achieve faster rates.

Our work has important points of contact with a recent line of papers on algorithms for non-convex and non-smooth problems, and we discuss a few of them here. Bolte et al. (2014) developed a proximal-type algorithm applicable to objective functions formed as a sum of smooth (possibly non-convex) and a convex (possibly non-differentiable) function. Some recent work by Xu and Yin (2017) extended these ideas and provided analysis for block co-ordinate descent methods for non-convex functions. Hong et al. (2016) analyzed the ADMM method for non-convex problems, whereas in other recent work (An and Nam, 2017; Wen et al., 2018), the authors proposed a proximal-type method for non-convex functions that can be written as a sum of a smooth function, a concave continuous function and a convex lower semi-continuous function; we also analyze this class in one of our results (Theorem 2).

Our results also relate to another interesting sub-area of non-convex optimization, namely functions that can be represented as a difference of two convex functions, popularly known as DC functions. We refer the reader to the papers (Tuy, 1995; Hartman, 1959; Lanckriet and Sriperumbudur, 2009; Yuille and Rangarajan, 2003) for more details on DC functions and their properties. One of the most popular DC optimization algorithms is the Convex Concave Procedure, or CCCP for short; see the papers (Yuille and Rangarajan, 2003; Lipp and Boyd, 2016) for further details. This is a double loop algorithm that minimizes a convex relaxation of the non-convex objective function at each iteration. While the CCCP algorithm has some attractive convergence properties (Lanckriet and Sriperumbudur, 2009), it can be slow in many situations due to its double loop structure. One outcome of the analysis in this paper is a single-loop proximal-method that retains all the convergence guarantees of CCCP while—as shown in our experimental results—being much faster to run.

1.1. Problem setup

In this paper, we study the problem of minimizing a non-convex and possibly non-smooth function over a closed convex set. More precisely, we consider optimization problems of the form

$$\min_{x\in\mathcal{C}} \Big\{ \underbrace{g(x) - h(x) + \varphi(x)}_{f(x)} \Big\}, \tag{1}$$

where the domain $\mathcal{C}$ is a closed convex set. In all cases, we assume the function f is bounded below over domain $\mathcal{C}$, and that the function h is continuous and convex. Our aim is to derive algorithms for problem (1) for various types of functions g and φ.

Structural assumption on functions g and h

(a) Theorems 1 and 4 are based on the assumption that the function g is continuously differentiable and smooth, and that the function $\varphi \equiv 0$.

(b) In Theorems 2 and 5, we assume that the function g is continuously differentiable and smooth, and that the function φ is convex, proper and lower semi-continuous.[1]

(c) Theorem 3 focuses on the case in which the function g is continuously differentiable, and the function $\varphi \equiv 0$.

The class of non-convex functions covered in part (a) includes, as a special case, the class of differences of convex (DC) functions, for which the first convex function is smooth and the second convex function is continuous. Note that we only put a mild assumption of continuity on the convex function h, meaning that the difference function $g - h$ can be non-smooth and non-differentiable in general. In particular, for any continuously differentiable function h and any smooth function g, the difference function $f = g - h$ is non-smooth. Furthermore, if we take the function $h \equiv 0$, then we recover the class of smooth functions as a special case.

1.2. Overview of our results

- Our first main result (Theorem 1) provides guarantees for a subgradient algorithm as applied to the minimization problem (2), to be defined in the sequel, when constrained to a closed convex set $\mathcal{C}$. We provide convergence bounds in terms of the Euclidean norm of

1. Taking the function $\varphi \equiv 0$ yields part (a) as a special case, but it is worthwhile to point out that the assumptions in Theorem 1 are weaker than the assumptions of Theorem 2. Furthermore, we can prove some interesting results about saddle points when the function $\varphi \equiv 0$; see Corollary 3.

the subgradient and show that our rates are unimprovable in general. We also illustrate some consequences of Theorem 1 by deriving a convergence rate for our algorithm when applied to non-smooth coercive functions; this result has interesting implications for polynomial programming. We also provide a simplification of the CCCP algorithm, along with convergence guarantees. In Corollary 3, we argue that our algorithm can escape strict saddle points for a large class of non-smooth functions, thereby generalizing known results for smooth functions.

- Our second main result (Theorem 2) provides convergence rates for a proximal-type algorithm for problem (1). In Section 4.3, we demonstrate how this proximal-type algorithm can be used to minimize a smooth convex function subject to a sparsity constraint. We demonstrate the performance of this algorithm through the example of best subset selection.
- In Theorem 3, we provide a Frank-Wolfe type algorithm for solving optimization problem (17), and we provide a rate of convergence in terms of the associated Frank-Wolfe gap.
- Finally, in Theorems 4 and 5, we prove that Algorithms 1 and 2, when applied to functions that satisfy a variant of the Kurdaya-Łojasiewicz inequality, have faster convergence rates. In particular, the convergence rate in terms of gradient norm is at least $\mathcal{O}(1/k)$ – whereas the worst case rate for general non-convex functions is $\mathcal{O}(\frac{1}{\sqrt{k}})$. We also provide examples of functions for which the convergence rate is $\mathcal{O}(1/k^r)$ with $r > 1$. In Theorem 6, we characterize the class of functions that can be written as a difference of a smooth function and a differentiable convex function.

Section 4 is devoted to an illustration of our methods and theory via applications to the problems of best subset selection, robust estimation, mixture density estimation and shape-from-shading reconstruction.

Notation: Given a set $\mathcal{C} \subset \mathbb{R}^d$, we use $\text{int}(\mathcal{C})$ to denote its interior. We use $\|x\|_2$, $\|x\|_1$ and $\|x\|_0$ to denote the Euclidean norm, ℓ_1-norm and ℓ_0 norms, respectively, of a vector $x \in \mathbb{R}^d$. We say that a continuously differentiable function g is M_g-smooth if the gradient ∇g is M_g-Lipschitz continuous. In many examples considered in this paper, the objective function f is a linear combination of a differentiable function g and one or more convex functions h and φ. With a slight abuse of notation, for a function $f = g - h + \varphi$, we refer to a vector of the form $\nabla g(x) - u(x) + v(x)$, where $u(x) \in \partial h(x)$ and $v(x) \in \partial \varphi(x)$, as a gradient of the function f at point x — and we denote it by $\nabla f(x)$; here, $\partial h(\cdot)$ and $\partial \varphi(\cdot)$ denote the subgradient sets of the convex functions h and φ respectively. We say a point x is a *critical* point of the function f if $0 \in \nabla f(x)$. For a sequence $\left\{a^k\right\}_{k \geq 0}$, we define the running arithmetic mean $\text{Avg}\left(a^k\right)$ as $\text{Avg}\left(a^k\right) := \frac{1}{k}\sum_{\ell=0}^{\ell=k+1} a^\ell$. Similarly, for a non-negative sequence $\left\{a^k\right\}_{k \geq 0}$, we use $\text{GAvg}\left(a^k\right) := (\prod_{\ell=0}^{k} a^\ell)^{\frac{1}{k+1}}$ to denote the running geometric mean. Finally, for real-valued sequences $\{a^k\}_{k \geq 0}$ and $\{b^k\}$, we say $a^k = \mathcal{O}\left(b^k\right)$, if there exists a positive constant C, which is independent of k, such that $a^k \leq C b^k$ for all $k \geq 0$. We say $a^k = \Omega(b^k)$ if $a^k = \mathcal{O}\left(b^k\right)$ and $b^k = \mathcal{O}\left(a^k\right)$.

2. Main results

Our main results are analyses of three algorithms for this class of non-convex non-smooth problems; in particular, we derive non-asymptotic bounds on their rates of convergence. The first algorithm is a (sub)-gradient-type method, and it is mainly suited for unconstrained optimization; the second algorithm is based on a proximal operator and can be applied to constrained optimization problems. The third algorithm is a Frank-Wolfe-type algorithm, which is also suitable for constrained optimization problems, but it applies to a more general class of non-convex optimization problems.

2.1. Gradient-type method

In this section, we analyze a (sub)-gradient-based method for solving a certain class of non-convex optimization problems. In particular, consider a pair of functions (g, h) such that:

Assumption GR:

(a) The function g is continuously differentiable and M_g-smooth.

(b) The function h is continuous and convex.

(c) There is a closed convex set $\mathcal{C}$ such that the difference function $f := g - h$ is bounded below on the set $\mathcal{C}$.

Under these conditions, we then analyze the behavior of a (sub)-gradient method in application to the following problem

$$f^* = \min_{x\in\mathcal{C}} f(x) \ = \ \min_{x\in\mathcal{C}} \big\{g(x) - h(x)\big\}. \tag{2}$$

Let $\partial h(x)$ denote the subdifferential of the convex function h at the point x. With a slight abuse of notation, we refer to a vector of the form $\nabla g(x) - u(x)$ with $u(x) \in \partial h(x)$ as a gradient of the function f at the point x.

Algorithm 1 Subgradient-type method

1: Given an initial point $x^0 \in \text{int}(\mathcal{C})$ and step size $\alpha \in (0, \frac{1}{M_g}]$:
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Choose subgradient $u^k \in \partial h(x^k)$.
4: Update $x^{k+1} = x^k - \alpha\big(\nabla g(x^k) - u^k\big)$.
5: **end for**

In our analysis, we assume that the initial vector $x^0 \in \text{int}(\mathcal{C})$ is chosen such that the associated level set

$$\mathcal{L}(f(x^0)) := \big\{x \in \mathbb{R}^d \mid f(x) \leq f(x^0)\big\}$$

is contained within $\text{int}(\mathcal{C})$. This condition is standard in the analysis of non-convex optimization methods (e.g., see Nesterov and Polyak 2006). When $\mathcal{C} = \mathbb{R}^d$, it holds trivially. With this set-up, we have the following guarantees on the convergence rate of Algorithm 1.

Theorem 1 *Under Assumption GR, any sequence $\{x^k\}_{k\geq 0}$ produced by Algorithm 1 has the following properties:*

(a) Any limit point is a critical point of the function f, and the sequence of function values $\{f(x^k)\}_{k\geq 0}$ is strictly decreasing and convergent.

(b) For all $k = 0, 1, 2, \ldots$, we have

$$\mathrm{Avg}\left(\|\nabla f(x^k)\|_2^2\right) \leq \frac{2\big(f(x^0) - f^*\big)}{\alpha(k+1)}. \tag{3}$$

See Appendix B.1 for a proof of this theorem.

2.1.1. Comments on convergence rates

Note that the bound (3) guarantees that the gradient norm sequence $\min_{j\leq k} \|\nabla f(x^j)\|_2$ converges to zero at the rate $\mathcal{O}(1/\sqrt{k})$. It is natural to wonder whether this convergence rate can be improved. Interestingly, the answer is no, at least for the general class of functions covered by Theorem 1. Indeed, note that the class of M-smooth functions is contained within the class of functions covered by Theorem 1. It follows from past work by Cartis et al. (2010) that for gradient descent on M-smooth functions, with a step size chosen according to the Goldstein-Armijo rule, the convergence rate of the gradient sequence $\{\|\nabla f(x^k)\|_2\}_{k\geq 0}$ can be lower bounded—for appropriate choices of the function f—as $\Omega(1/\sqrt{k})$. It is not very difficult to see that the same construction also provides a lower bound of $\Omega(1/\sqrt{k})$ for gradient descent with a constant step size. We also note that very recently, Carmon et al. (2017) proved an even stronger result: more precisely, for the class of smooth functions, the rate of convergence of any algorithm given access to only the function gradients and function values cannot be faster than $\Omega(1/\sqrt{k})$. Finally, observe that in the special case $h \equiv 0$, Algorithm 1 reduces to the ordinary gradient descent with fixed step size α. Putting together the pieces, we conclude that for the class of functions which can be written as a difference of smooth and a continuous convex function, Algorithm 1 is *optimal* among all algorithms that have access to the gradients (and/or the sub-gradients) and the function values.

2.2. Consequences for differentiable functions

In the special case when the function h is convex and differentiable, Algorithm 1 reduces to an ordinary gradient descent on the difference function $f = g - h$. However, note that the step size choice required in Algorithm 1 does *not* depend on the smoothness of the function h; consequently, the algorithm can be applied to objective functions f that are not smooth. As a simple but concrete example, suppose that we wish to apply gradient descent to minimize the function $f(x) := g(x) - \|x\|_2^q$, where g is any μ-strongly convex and M_g-smooth function, and $q \in (1, 2)$ is a given parameter. Classical guarantees on gradient descent, which require the smoothness of the function f, would not apply here since the function f itself is not smooth. However, Theorem 1 guarantees that standard gradient descent would converge for any step size $\alpha \in \big(0, \frac{1}{M_g}\big]$.

More generally, given an arbitrary continuously differentiable function f, we can define its *effective smoothness constant* as

$$M_f^* := \inf_h \big\{L \mid (f+h) \text{ is } L\text{-smooth}\big\}, \tag{4}$$

where the infimum ranges over all convex and continuously differentiable functions h. Suppose that this infimum is achieved by some function h^*, then gradient descent on the function f can be viewed as applying Algorithm 1 to the decomposition $f = g^* - h^*$, where the function $g^* := f + h^*$ is guaranteed to be M_f^*-smooth. To be clear, the algorithm itself does *not* need to know the decomposition (g^*, h^*), but the existence of the decomposition ensures the success of a backtracking procedure. Putting together the pieces, we arrive at the following consequence of Theorem 1:

Corollary 1 *Given a closed convex set $\mathcal{C}$, consider a continuously differentiable function f with effective smoothness $M_f^* < \infty$ that is bounded below on $\mathcal{C}$. Then for any sequence $\{x^k\}_{k\geq 0}$ obtained by applying the gradient update with step size $\alpha \in \big(0, \frac{1}{M_f^*}\big)$, we have:*

$$\operatorname{Avg}\Big(\|\nabla f(x^k)\|_2^2\Big) \leq \frac{2\big(f(x^0) - f^*\big)}{\alpha(k+1)}. \tag{5a}$$

Moreover, if we choose step size by backtracking[2] *with parameter $\beta \in (0,1)$, then for all $k = 0, 1, 2, \ldots$, we have*

$$\operatorname{Avg}\Big(\|\nabla f(x^k)\|_2^2\Big) \leq \frac{2\max\big\{1, M_f^*\big\}\big(f(x^0) - f^*\big)}{\beta^2(k+1)}. \tag{5b}$$

See Appendix B.2 for proof of the above corollary.

Let us reiterate that the advantage of backtracking gradient descent is that it works without knowledge of the scalar M_f^*. The parameter β mentioned in equation (5b) is the user-defined backtracking parameter (see Algorithm 4 for details). In particular, substituting $\beta = \frac{1}{\sqrt{2}}$ in equation (5b) yields

$$\operatorname{Avg}\Big(\|\nabla f(x^k)\|_2^2\Big) \leq \frac{4\max\big\{1, M_f^*\big\}\big(f(x^0) - f^*\big)}{(k+1)},$$

which differs from the rate obtained in equation (5a) only by a factor of two, and a possible multiple of M_f^*.

2.2.1. Consequences for coercive functions

As a consequence of Corollary 1, we can obtain a rate of convergence of the backtracking gradient descent algorithm (Algorithm 4) for a class of non-smooth coercive functions. Consider any twice continuously differentiable coercive function $f : \mathbb{R}^d \mapsto \mathbb{R}$, which is bounded below. Recall that a function f is *coercive* if

$$f(x^\ell) \overset{\ell\to\infty}{\to} \infty \quad \text{for any sequence } \{x^\ell\}_{\ell\geq 0} \text{ such that } \quad \|x^\ell\|_2 \to \infty. \tag{6}$$

2. A detailed description of gradient descent with backtracking is provided in Algorithm 4.

Let $\mathcal{L}(f(x^0)) := \{x \in \mathbb{R}^d : f(x) \leq f(x^0)\}$ denote the level set of the function f at point x^0. It can be verified that for any coercive function f, the set $\mathcal{L}(f(x^0))$ is bounded above for all $x^0 \in \mathbb{R}^d$. This property ensures that for any descent algorithm and any starting point x^0, the set of iterates $\{x^k\}_{k \geq 0}$ obtained from the algorithm remains within a bounded set—viz. the level set $\mathcal{L}(f(x^0))$ in this case. Since the function f is twice continuously differentiable, we have that f is smooth over bounded set $\mathcal{L}(f(x^0))$; this fact ensures that f has a finite effective smoothness constant in the set $\mathcal{L}(f(x^0))$, which we denote by M^*_{f,x^0}. Finally, note that Algorithm 4 is a descent algorithm; as a result, a simple application of Corollary 1 yields the following rate of convergence for the backtracking gradient descent algorithm (Algorithm 4):

Corollary 2 *Consider the unconstrained minimization problem of a twice continuously differentiable coercive function f that is bounded below on $\mathbb{R}^d$. Then for any initial point x^0, the sequence $\{x^k\}_{k \geq 0}$ obtained by applying Algorithm 4 satisfies the following property:*

$$\text{Avg}\left(\|\nabla f(x^k)\|_2^2\right) \leq \frac{2 \max\{1, M^*_{f,x^0}\}\left(f(x^0) - f^*\right)}{\beta^2(k+1)} \quad \textit{for all } k = 0, 1, 2, \ldots, \tag{7}$$

where $\beta \in (0, 1)$ is the backtracking parameter.

Implications for polynomial programming: Corollary 2 has useful implications for problems that involve minimizing polynomials. Such problems of polynomial programming arise in various applications, including phase retrieval and shape-from-shading (Wang et al., 2014), and we illustrate our algorithms for the latter application in Section 4.1. For minimization of a coercive polynomial, Corollary 2 shows that Algorithm 4 achieves a near-optimal rate.

It is worth noting that any even degree polynomial can be represented as a difference of convex (DC) function; hence, such problems are amenable to DC optimization techniques like CCCP, which we discuss at more length in Section 2.3. However, obtaining a good DC decomposition, which is crucial to the success of CCCP, is often a formidable task. In particular, obtaining an optimal decomposition for a polynomial with degree greater than four is NP-hard; indeed, deciding the convexity of an even degree polynomial with degree greater than four is NP-hard (Ahmadi et al., 2013; Wang et al., 2014). Even for a fourth degree polynomial with dimension larger than three, there is no known algorithm for finding an optimal DC decomposition (Ahmadi and Parrilo, 2013). An advantage of Algorithm 4 is that it obviates the need to find a DC decomposition.

2.2.2. Escaping strict saddle points

One of the obstacles with gradient-based continuous optimization method is possible convergence to saddle points. Here we show that with a random initialization this undesirable outcome does not occur for the class of strict saddle points. Recall that for a twice differentiable function f, a point x is called a strict saddle point of the function f if $\lambda_{\min}(\nabla^2 f(x)) < 0$, where $\lambda_{\min}(\nabla^2 f(x))$ denotes the minimum eigenvalue of the Hessian matrix $\nabla^2 f(x)$. The following corollary shows that such saddle points are *not* troublesome:

Corollary 3 *Suppose that, in addition to the conditions on $(g, h, \mathcal{C})$ from Theorem 1, the functions (g, h) are twice continuously differentiable. If Algorithm 1 is applied with step size*

$\alpha \in \left(0, \frac{1}{M_g}\right)$, *then the set of initial points for which it converges to a strict saddle point has measure zero.*

See Appendix B.3 for the proof of this corollary.

We note that similar guarantees of avoidance of strict saddle-points are known when the function $f = g - h$ is twice continuously differentiable and M-smooth (e.g., Lee et al. 2016; Panageas and Piliouras 2016). The novelty of Corollary 3 is that the same guarantee holds without imposing a smoothness condition on the entire function f.

2.3. Connections to the convex-concave procedure

As a consequence of Algorithm 1, we show that one can obtain a convergence rate of the Euclidean norm of the gradient for CCCP (convex-concave procedure), which is a heavily used algorithm in Difference of Convex (DC) optimization problems. Before doing so, let us provide a brief description of DC functions and the CCCP algorithm.

DC functions: Given a convex set $\mathcal{C} \subseteq \mathbb{R}^d$, we say that a function $f : \mathcal{C} \mapsto \mathbb{R}$ is DC if there exist convex functions g and h with domain $\mathcal{C}$ such that $f = g - h$. Note that the DC representation $f = g - h$ mentioned in the definition is not unique. In particular, for any convex function p, we can write $f = (g + p) - (h + p)$. The class of DC functions includes a large number of non-convex problems encountered in practice. Both convex and concave functions are DC in a trivial sense, and the class of DC functions remains closed under addition and subtraction. More interestingly, under mild restrictions on the domain, the class of non-zero DC functions is also closed under multiplication, division, and composition (e.g., Tuy 1995; Hartman 1959). The maximum and minimum of a finite collection of DC functions are also DC functions.

Convex-concave procedure: An interesting class of problems are those that involve minimizing a DC function over a closed convex set $\mathcal{C} \subseteq \mathbb{R}^d$, i.e.

$$f^* := \min_{x \in \mathcal{C}} f(x) \; = \; \min_{x \in \mathcal{C}} \big\{ g(x) - h(x) \big\}, \tag{8}$$

where g and h are proper convex functions. The above problem has been studied intensively, and there are various methods for solving it; for instance, see the papers (Tuy, 1995; Lipp and Boyd, 2016; Pham Dinh et al., 2013) and references therein for details. One of the most popular algorithms to solve problem (8) is the Convex-concave Procedure (CCCP), which was introduced by Yuille and Rangarajan (2003). The CCCP algorithm is a special case of a Majorization-Minimization algorithm, one which uses the DC structure of the objective function in problem (8) to construct a convex majorant of the objective function f at each step. We start with a feasible point $x^0 \in \mathrm{int}(\mathcal{C})$. Let x^k denote the iterate at k^{th} iteration; at the $(k+1)^{\text{th}}$ iteration we construct a convex majorant $q(\cdot,\, x^k)$ of the function f via

$$f(x) \leq \underbrace{g(x) - h(x^k) - \langle u^k,\, x - x^k \rangle}_{=:\, q(x, x^k)}, \tag{9}$$

where $u^k \in \partial h(x^k)$, the subgradient set of the convex function h at point x^k. The next iterate x^{k+1} is obtained by solving the convex program

$$x^{k+1} \in \arg\min_{x \in \mathcal{C}} q(x, x^k). \tag{10}$$

The CCCP algorithm has some attractive convergence properties. For instance, it is a descent algorithm; when the function g is strongly convex differentiable and the function h is continuously differentiable, it can be shown (Lanckriet and Sriperumbudur, 2009) that any limit point of the sequence $\{x^k\}_{k\geq 0}$ obtained from CCCP is stationary. Under the same assumptions, one can also verify that $\lim_{k\to\infty} \|x^k - x^{k+1}\|_2 = 0$.

We now turn to an analysis of CCCP using the techniques that underlie Theorem 1. In the next proposition, we derive a rate of convergence of the gradient sequence and show that all limit points of the sequence $\{x^k\}_{k\geq 0}$ are stationary. Earlier analyses of CCCP, including the papers (Lanckriet and Sriperumbudur, 2009; Yuille and Rangarajan, 2003), are mainly based on the assumption of strong convexity of the function g, whereas in the next proposition, we only assume that the function g is M_g-smooth. When the function g is strongly convex, our analysis recovers the well-known convergence result in past work (Lanckriet and Sriperumbudur, 2009). In particular, we show that CCCP enjoys the same rate of convergence as that of Algorithm 1.

Proposition 1 *Under Assumption GR and with the function g being convex, the CCCP sequence* (10) *has the following properties:*

(a) Any limit point of the sequence $\{x^k\}_{k\geq 0}$ is a critical point, and the sequence of function values $\{f(x^k)\}_{k\geq 0}$ is strictly decreasing and convergent.

(b) Furthermore, for all $k = 1, 2, \ldots$, we have

$$\operatorname{Avg}\Big(\|\nabla f(x^k)\|_2^2\Big) \leq \frac{2M_g\big(f(x^0) - f^*\big)}{(k+1)}, \tag{11a}$$

and assuming moreover that g is μ-strongly convex,

$$\operatorname{Avg}\Big(\|x^k - x^{k+1}\|_2^2\Big) \leq \frac{2\big(f(x^0) - f^*\big)}{\mu(k+1)}. \tag{11b}$$

The proof of this proposition builds on the argument used for Theorem 1; see Appendix B.4 for details.

2.3.1. Simplifying CCCP

Algorithm 1 provides us an alternative procedure for minimizing a difference of convex functions when the first convex function is smooth. The benefit of Algorithm 1 over standard CCCP is that Algorithm 1 is a single loop algorithm and is expected to be faster than standard double loop CCCP algorithm in many situations. Furthermore, Algorithm 1 shares convergence guarantees similar to a standard CCCP algorithm.

2.4. Proximal-type method

We now turn to a more general class of optimization problems of the form

$$f^* := \min_{x\in\mathbb{R}^d} f(x) \;=\; \min_{x\in\mathbb{R}^d} \Big\{ \big(g(x) - h(x)\big) + \varphi(x) \Big\}. \tag{12}$$

We assume that the functions g, h and φ satisfy the following conditions:

Assumption PR

(a) The function $f = g - h + \varphi$ is bounded below on $\mathbb{R}^d$.

(b) The function g is continuously differentiable and M_g-smooth; the function h is continuous and convex; and the function φ is proper, convex and lower semi-continuous.

Typical examples of the function φ include $\varphi(x) = \|x\|_1$, or the indicator of a closed convex convex set $\mathcal{X}$. Since for a general lower semi-continuous function φ, the sum-function $g + \varphi$ is neither differentiable nor smooth, a gradient-based method cannot be applied. One way to minimize such functions is via a proximal-type algorithm, of which the following is an instance.

Algorithm 2 Proximal-type algorithm

1: Given an initial vector $x^0 \in \mathrm{dom}(f)$ and step size $\alpha \in \big(0, \frac{1}{M_g}\big]$.
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Update $x^{k+1} = \mathrm{prox}^{\varphi}_{1/\alpha}\Big(x^k - \alpha\big(\nabla g(x^k) - u^k\big)\Big)$ for some $u^k \in \partial h(x^k)$.
4: **end for**

The proximal update in line 3 of Algorithm 2 is very easy to compute and often has a closed form solution (see (Parikh et al., 2014)). Let us now derive the rate of convergence result of Algorithm 2.

Theorem 2 *Under Assumption PR, any sequence $\{x^k\}_{k\geq 0}$ obtained from Algorithm 2 has the following properties:*

(a) Any limit point of the sequence $\{x^k\}_{k\geq 0}$ is a critical point, and the sequence of function values $\{f(x^k)\}_{k\geq 0}$ is strictly decreasing and convergent.

(b) For all $k = 1, 2, \ldots$, we have

$$\mathrm{Avg}\Big(\|x^k - x^{k-1}\|_2^2\Big) \leq \frac{2\alpha\big(f(x^0) - f^*\big)}{(k+1)}. \tag{13a}$$

If moreover the function h is M_h-smooth, then

$$\mathrm{Avg}\Big(\|\nabla f(x^k)\|_2^2\Big) \leq \frac{2\alpha C_{M,\alpha}\big(f(x^0) - f^*\big)}{(k+1)}, \tag{13b}$$

where $C_{M,\alpha} = \big(M_g + M_h + \frac{1}{\alpha}\big)^2$.

See Appendix C for the proof of the theorem.

Comments: The proof of Theorem 2 reveals that the smoothness condition on the function h in Theorem 2 can be replaced by the local smoothness of h, when the sequence $\{x^k\}_{k\geq 0}$ is bounded. Note that the local smoothness condition is weaker than the global smoothness condition. For instance, any twice continuously differentiable function is locally smooth. The boundedness assumption on the iterates $\{x^k\}_{k\geq 0}$ holds in many situations. For instance, if the function f is coercive (6), then it follows that the iterates $\{x^k\}_{k\geq 0}$ remain bounded. Another instance is when the function φ is the indicator function of a compact convex set. Finally, we point out that when the function h is non-smooth but the proximal-function φ is smooth, the existing proof can be easily modified to obtain a rate of convergence of the gradient-norm $\|\nabla f(x^k)\|_2$.

Projected Gradient Descent: A special case of the Algorithm 2 is when φ is equal to the indicator function $\mathbb{1}_{\mathcal{X}}$ of a closed convex set $\mathcal{X}$. Consider the following constrained optimization problem

$$f^* := \min_{x\in\mathcal{X}} \; \{\underbrace{g(x) - h(x)}_{f(x)}\}, \tag{14}$$

where $\mathcal{X}$ is a closed convex set, the function g is M_g-smooth, and the function h is convex continuous. Using Algorithm 2, the update equation in this case is given by

$$x^{k+1} = \Pi_{\mathcal{X}}\big(x^k - \alpha(\nabla g(x^k) - u^k)\big). \tag{15}$$

In projected-gradient-type methods, we should not expect a rate in terms of the gradient. In such cases, the projected gradient step may not be aligned with the gradient direction, or the step size may be arbitrarily small due to projection. Rather, an appropriate analogue of the gradient in this case is as follows:

$$\nabla f_{\mathcal{X}}(x^k) = \frac{1}{\alpha}\big(x^k - \Pi_{\mathcal{X}}(x^k - \alpha(\nabla g(x^k) - u^k))\big). \tag{16}$$

The analysis of the projected gradient method using $\nabla f_{\mathcal{X}}(x^k)$ is standard in the optimization literature (Bubeck et al., 2015). It is worth pointing out that the quantity $\nabla f_{\mathcal{X}}(x^k)$ is the analogue of the gradient in the constrained optimization setup, and coincides with the gradient in the unconstrained setup. Concretely, we have $\nabla f_{\mathcal{X}}(x^k) = \nabla f(x^k)$ where $f := g - h$, and $\mathcal{X} = \mathbb{R}^d$. Combining equations (15) and (16) and applying the bound (13b) from Theorem 2, we find that

$$\text{Avg}\left(\|\nabla f_{\mathcal{X}}(x^k)\|_2^2\right) \leq \frac{2\big(f(x^0) - f^*\big)}{\alpha(k+1)}.$$

2.5. Frank-Wolfe type method

In our analysis of the previous two algorithms, we assumed that the objective function f has a smooth component g, and we leveraged the smoothness property of g to establish convergence rates. In many situations, the objective function may not have a smooth component; consequently, neither the gradient-type algorithm nor the prox-type algorithm provides any theoretical guarantee. In this section, we analyze a Frank-Wolfe-type algorithm for solving such optimization problems. In particular, consider an optimization problem of the form

$$f^* := \min_{x \in \mathcal{C}} f(x) \;= \min_{x \in \mathcal{C}} \big\{ g(x) - h(x) \big\}, \tag{17}$$

where $\mathcal{C}$ is a closed convex set, and the functions (g, h) satisfy the following conditions:

Assumption FW:

(a) The difference function $f = g - h$ is bounded below over range $\mathcal{C}$.

(b) The function g is continuously differentiable, whereas the function h is convex and continuous.

The analysis of the Frank-Wolfe algorithm for a convex problem is based on the *curvature constant* $\mathcal{C}_f$ of the convex objective function with respect to the closed convex set $\mathcal{C}$. This curvature constant can be defined for any differentiable function, which need not be convex (Lacoste-Julien, 2016).

Here we define a slight generalization of this notion, applicable to a non-differentiable function $f = g - h$ that can be written as a difference of a differentiable function g and a continuous convex function h (which may be non-differentiable). Define the set

$$S_\gamma := \big\{ x, y \in \mathcal{C} \mid \text{there exist } \gamma \in (0,1] \text{ and } u \in \mathcal{C} \text{ with } y = x + \gamma(u - x) \big\},$$

and the curvature constant

$$\mathcal{C}_f = \sup_{\substack{x,y \in S_\gamma \\ u \in \partial h(x)}} \frac{2}{\gamma^2} \big[f(y) - f(x) - \langle y - x,\, \nabla g(x) - u \rangle \big]. \tag{18}$$

Note that in the special case $h \equiv 0$, we recover the curvature constant of the differentiable function g used by Lacoste-Julien (Lacoste-Julien, 2016). We refer to the scalar $\mathcal{C}_f$ as the generalized curvature constant of the function f with respect to the closed convex set $\mathcal{C}$.

Algorithm 3 Frank-Wolfe type method

1: Given initial vector $x^0 \in \int(\mathcal{C})$:
2: **for** $k = 1, \ldots, K$ **do**
3: Choose any $u^k \in \partial h(x^k)$.
4: Compute $s^k := \arg\min_{s \in \mathcal{C}} \langle s,\, \nabla g(x^k) - u^k \rangle$.
5: Define $d^k := s^k - x^k$ and $g^k := -\langle d^k,\, \nabla g(x^k) - u^k \rangle$. *(Frank-Wolfe gap)*
6: Set $\gamma^k = \min \big\{ \frac{g^k}{C_0}, 1 \big\}$ for some $C_0 \geq \mathcal{C}_f$.
7: Update $x^{k+1} = x^k + \gamma^k d^k$.
8: **end for**

Next, we provide an analysis of Algorithm 3 in terms of the Frank-Wolfe (FW) gap g^k defined Step 5. We show that the minimum FW gap $\{g^k\}_{k \geq 0}$ defined in Algorithm 3 converges to zero at the rate $\frac{1}{\sqrt{k+1}}$.

Theorem 3 *Under Assumption FW, the Frank-Wolfe gap sequence* $\{g^k\}_{k\geq 0}$ *from Algorithm 3 satisfies the following property:*

$$\min_{0\leq j\leq k} g^j \leq \frac{\max\left\{2\big(f(x^0)-f^*\big), C_0\right\}}{\sqrt{k+1}} \qquad \text{for all } k = 0, 1, 2, \ldots.$$

See Appendix D.1 for the proof of this theorem.

Comments: The FW gap appearing in Theorem 3 is standard in the analysis of Frank-Wolfe algorithm; note that it is invariant to an affine transformation of the set $\mathcal{C}$. Similar convergence guarantees for the minimum FW-gap are available for differentiable functions; for instance, see the paper (Lacoste-Julien, 2016). The novelty of the above theorem is that it provides convergence guarantees of minimum FW-gap for a class of non-differentiable functions.

Upper bound on generalized curvature constant: It is worth mentioning that Algorithm 3 only requires an upper bound of the generalized curvature constant $\mathcal{C}_{g-h}$. Consequently, it is interesting to obtain an upper bound for the scalar $\mathcal{C}_{g-h}$. For a M_g-smooth function g, one well-known upper bound of the curvature constant $\mathcal{C}_g$ is $M_g \times \big(\operatorname{diam}_{\|\cdot\|_2}(\mathcal{C})\big)^2$; see also Jaggi (2013). A similar upper bound also holds for the generalized curvature constant defined in equation (59). In particular, we prove that for a difference function $f = g-h$, with the function h being convex continuous, the scalar $\mathcal{C}_{g-h}$ is always upper bounded by $\mathcal{C}_g$, the curvature constant of the function g (see Lemma 6).

3. Faster rate under KL-inequality

In the preceding sections, we have derived rates of convergence for the gradient norms for various classes of problems. It is natural to wonder if faster convergence rates are possible when the objective function is equipped with some additional structure. Based on Theorems 1 and 2, we see that both Algorithms 1 and 2 ensure that $\|x^k - x^{k+1}\|_2 \to 0$, meaning that the successive differences between the iterates converge to zero. Although we proved that any limit point of the sequence $\{x^k\}_{k\geq 0}$ has desirable properties, the condition $\|x^k - x^{k+1}\|_2 \to 0$ is not sufficient—at least in general—to prove convergence[3] of the sequence $\{x^k\}_{k\geq 0}$. In this section, we provide a sufficient condition under which Algorithm 1 and Algorithm 2 yield convergent sequences of iterates $\{x^k\}_{k\geq 0}$, and we establish that the gradient sequences $\{\|\nabla f(x)\|_2\}_{k\geq 0}$ converge at faster rates.

3.1. Kurdaya-Łojasiewicz inequality

Let us now establish a faster local rate of convergence of Algorithms 1 and 2 for functions that satisfy a form of the Kurdaya-Łojasiewicz (KL) inequality. More precisely, suppose that there exists a constant $\theta \in [0, 1)$ such that the ratio $\frac{(f(x)-f(\bar{x}))^\theta}{\|\nabla f(x)\|_2}$ is bounded above in a neighborhood of every point $\bar{x} \in \operatorname{dom}(f)$. This type of inequality is known as a Kurdaya-Łojasiewicz inequality, and the exponent θ is known as the Kurdaya-Łojasiewicz

3. The convergence of the sequence $\{x^k\}_{k\geq 0}$ for Algorithm 2 was studied in the papers (An and Nam, 2017; Wen et al., 2018). We provide the proof under a weaker set of assumptions.

exponent (*KL-exponent*) of the function f at the point $\bar{x}$. These type of inequalities were first proved by Łojasiewicz (1963) for real analytic functions; in later work, Kurdyka (1998) and Bolte et al. (2007) proved similar inequalities for non-smooth functions, and the authors also provided examples of many functions that satisfy a form of the KL inequality. See Appendix A.2 for further details on functions of the KL type.

Assumption KL: For any point[4] $\bar{x} \in \text{dom}(f)$, there exists a scalar $\theta \in [0, 1)$ such that the ratio $\frac{|f(x)-f(\bar{x})|^\theta}{\|\nabla f(x)\|_2}$ is bounded above in a neighborhood of $\bar{x}$.

3.2. Convergence guarantees

Theorem 4 *Under Assumptions GR and KL, any bounded sequence* $\{x^k\}_{k\geq 0}$ *obtained from Algorithm 1 satisfies the following properties:*

(a) The sequence $\{x^k\}_{k\geq 0}$ *converges to a critical point* $\bar{x}$*, and for all* $k = 1, 2, \ldots$

$$\text{Avg}\left(\|\nabla f(x^k)\|_2\right) \leq \frac{c_1}{k},$$

(b) Suppose that at the point $\bar{x}$*, the function* f *has a KL exponent* $\bar{\theta} \in \left[\frac{1}{2}, \frac{r}{2r-1}\right)$ *for some* $r > 1$*. Then we have*

$$\text{GAvg}\left(\|\nabla f(x^k)\|_2\right) \leq \frac{c_2}{k^r} \qquad \textit{for all} \quad k = 1, 2, \ldots,$$

where the constants (c_1, c_2) *are independent of* k*, but they may depend on the KL parameters at the point* $\bar{x}$*.*

See Appendix E.1 for proof of this theorem.

Comments: It is worth pointing out that Theorem 4 does *not* require the function h to satisfy any smoothness assumption. Such conditions are needed for applying Algorithm 2, so that Theorem 4 is based on milder conditions than Theorem 5.

Our next result is to exhibit a faster convergence rate for Algorithm 2 under the KL assumption:

Theorem 5 *Suppose that, in addition to Assumptions PR & KL, the function* h *in Algorithm 2 is locally smooth. Then any bounded sequence* $\{x^k\}_{k\geq 0}$ *obtained from Algorithm 2 satisfy the following properties:*

(a) The sequence $\{x^k\}_{k\geq 0}$ *converges to a critical point* $\bar{x}$*, and for all* $k = 1, 2 \ldots$

$$\text{Avg}\left(\|\nabla f(x^k)\|_2\right) \leq \frac{c_1}{k}.$$

4. It can be shown that such an inequality would hold at non-critical point of a continuous function f; see Remark 3.2 of (Bolte et al., 2007). Note that the parameter θ and the neighborhood mentioned in Assumption KL above may depend on the point $\bar{x}$.

(b) *Given some* $r > 1$, *suppose that at the point* $\bar{x}$ *the function* f *has a KL exponent* $\bar{\theta} \in \left[\frac{1}{2}, \frac{r}{2r-1}\right)$. *Then*

$$\mathrm{GAvg}\left(\|\nabla f(x^k)\|_2\right) \leq \frac{c_2}{k^r} \qquad \textit{for all} \quad k = 1, 2, \ldots,$$

where the constants (c_1, c_2) *are independent of* k, *but they may depend on the KL parameters at the point* $\bar{x}$.

See Appendix E.2 for the proof of this theorem.

Comments: Note that $\min_{1 \leq i \leq k} \|\nabla f(x^k)\|_2$ is upper bounded by the quantities $\mathrm{Avg}\left(\|\nabla f(x^k)\|_2\right)$ and $\mathrm{GAvg}\left(\|\nabla f(x^k)\|_2\right)$. It thus follows that the sequence $\{\|\nabla f(x^k)\|_2\}_{k \geq 0}$ converges to zero at a rate of at least $1/k$, thereby improving the rate of convergence of $\|\nabla f(x)\|_2$ obtained in Theorems 1 and 2. When $\theta < \frac{1}{2}$, a simple modification of the proof (using $\gamma = 2$) shows that, Algorithms 1 and 2 converge in a finite number of steps. Finally, we point out that when the function h is non-smooth but the proximal-function φ is smooth, the existing proof can be easily modified to obtain a rate of convergence of the gradient-norm $\|\nabla f(x^k)\|_2$.

4. Some illustrative applications

In this section, we study four interesting classes of non-convex problems that fall within the framework of this paper. We also discuss various consequences of Theorems 1—5 as well as Corollaries 1—3 when applied to these problems.

4.1. Shape from shading

The problem of shape from shading is to reconstruct the three-dimensional (3D) shape of an object based on observing a two-dimensional (2D) image of intensities, along with some information about the light source direction. It is assumed that the observed 2D image intensity is determined by the angle between the light source direction and the surface normals of the object (Ecker and Jepson, 2010).

In more detail, suppose that both the object and its 2D image are supported on a rectangular grid of size $r \times c$. We introduce the shorthand notation $[r] = \{1, 2, \ldots, r\}$ and $[c] = \{1, 2, \ldots, c\}$ for the rows and columns of this grid. For each pair $(i, j) \in [r] \times [c]$, we let $I_{ij} \in \mathbb{R}$ denote the observed intensity at location (i, j) in the image, and we let $N_{ij} \in \mathbb{R}^3$ denote the surface normal at the vertex $v_{ij} := (x_{ij}, y_{ij}, z_{ij})$ of the object. Based on observing the 2-dimensional image, both the intensity I_{ij} and co-ordinate pair (x_{ij}, y_{ij}) are known for each pair $(i, j) \in [r] \times [c]$. The goal of shape from shading is to estimate the unknown coordinate z_{ij}, which corresponds to the height of the object at location (i, j). Knowledge of these z-coordinates allows us to generate a 3D representation of the object, as illustrated in Figure 1.

Lambertian lighting model: In order to reconstruct the z-coordinates, we require a model that relates the observed intensity I_{ij} to the surface normal. In a Lambertian model, for a given light source direction $L := (\ell_1, \ell_2, \ell_2)^\top \in \mathbb{R}^3$, it is assumed that the surface normal N_{ij} and intensity I_{ij} are related via the relation

$$I_{ij} = \frac{\langle L,\, N_{ij}\rangle}{\|N_{ij}\|_2}. \tag{19}$$

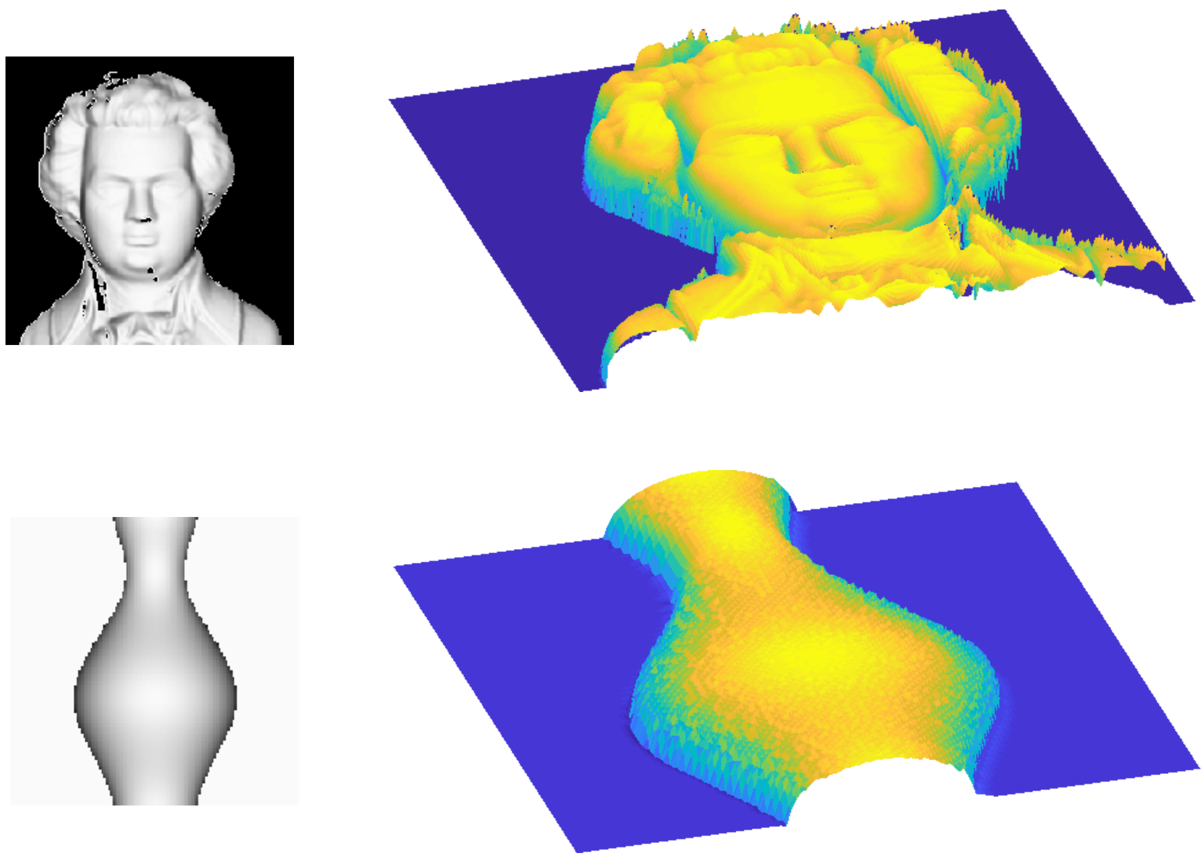

Figure 1. Figure shows 3D shape reconstruction of *Mozart* (first row) and *Vase* (second row) from corresponding 2D images. The gray-scale images in the left column are the 2D input images; the two colored images in the right column are the reconstructed 3D shapes. The 3D shapes are constructed by solving the problem (21) using Algorithm 4.

In one standard model (Wang et al., 2014), the surface normal $N_{ij} := (p_{ij}, q_{ij}, 1)^\top$ is assumed to be determined by the triplet of vertices $(v_{ij}, v_{i+1,j}, v_{i,j+1})$ via the equations

$$p_{ij} = \frac{(y_{i,j+1} - y_{i,j})(z_{i+1,j} - z_{ij}) - (y_{i+1,j} - y_{i,j})(z_{i,j+1} - z_{ij})}{(x_{i,j+1} - x_{ij})(y_{i+1,j} - y_{ij}) - (x_{i+1,j} - x_{ij})(y_{i,j+1} - y_{ij})},$$
$$q_{ij} = \frac{(x_{i,j+1} - x_{i,j})(z_{i+1,j} - z_{ij}) - (x_{i+1,j} - x_{i,j})(z_{i,j+1} - z_{ij})}{(x_{i,j+1} - x_{ij})(y_{i+1,j} - y_{ij}) - (x_{i+1,j} - x_{ij})(y_{i,j+1} - y_{ij})}.$$

Squaring both sides of equation (19) and substituting the expression for surface normal N_{ij} yields the polynomial equation

$$\big(p_{ij}^2 + q_{ij}^2 + 1\big)I_{ij} - (\ell_1 p_{ij} + \ell_2 q_{ij} + \ell_3)^2 = 0, \tag{20}$$

which should be satisfied under the assumed model.

In practice, this equality will not be exactly satisfied, but we can estimate the z-coordinates by solving the following non-convex optimization problem in the $r \times c$ matrix z with entries $\big\{z_{ij} \mid (i,j) \in [r] \times [c]\big\}$:

$$\min_{z \in \mathbb{R}^{r \times c}} \quad \Big\{ \underbrace{\sum_{i=1}^{r} \sum_{j=1}^{c} \big((1 + p_{ij}^2 + q_{ij}^2) I_{ij}^2 - (\ell_1 p_{ij} + \ell_2 q_{ij} + \ell_3)^2\big)^2}_{P(z)} \Big\}. \tag{21}$$

Some reconstruction experiments: In order to illustrate the behavior of our method for this problem, we considered two synthetic images for simulated experiments. The first one is a 256×256 image of *Mozart* (Zhang et al., 1999), and the second one is a 128×128 image of *Vase*. The 3D shapes were constructed from the 2D images by solving optimization problem (21) using the backtracking gradient descent algorithm 4. The reconstructed surfaces for *Vase* and *Mozart* are provided in Figure 1. We ran 500 iterations of Algorithm 4 for both the images. The runtime for *Mozart*-example was 87 seconds, whereas the runtime for *Vase*-example was 39 seconds. The implementation of Algorithm 4 for Problem (21) is parallelizable; hence, the runtime can be much lower than our runtime with a parallel implementation. It is worth mentioning that the polynomial P is a fourth-degree polynomial with dimension $r \times c$; polynomial P is coercive and bounded below by zero. Consequently, we can apply Corollary 2 to the problem (21) which guarantees that average of the squared gradient norm $\text{Avg}\left(\|\nabla P\|_2^2\right)$ converges to zero at a rate $\frac{1}{k}$.

One might also consider applying the CCCP method to this problem. In a recent paper, Wang et al. (2014) provided a DC decomposition of the polynomial P using a sum of square (SOS) optimization technique. However, it is crucial to note that the DC decomposition of polynomial P obtained from the SOS-optimization method need not be optimal. In order to see this, note that the dimension of the polynomial P is much larger than three. In particular, the variable z_{ij} is used in the computation of surface normals $N_{ij}, N_{i,j-1}$ and $N_{i-1,j}$, hence is related to variables $(z_{i,j+1}, z_{i+1,j}, z_{i-1,j}, z_{i,j-1})$—which are again related to the other variables. Ahmadi and Parrilo (2013) showed that SOS techniques for deriving a DC decomposition are sub-optimal for a fourth-degree polynomial when the dimension of the polynomial is greater than three. Consequently, deriving an optimal DC decomposition for the polynomial P will be computationally intensive.

4.2. Robust regression using Tukey's bi-weight

Next, we turn to the problem of robust regression with Tukey's bi-weight penalty function. Suppose that we observe pairs $(y_i, z_i) \in \mathbb{R} \times \mathbb{R}^d$ linked via the noisy linear model

$$y_i = \langle z_i,\, \mu^* \rangle + \varepsilon_i \qquad \text{for } i = 1, \ldots, n.$$

Here the vector $\mu^* \in \mathbb{R}^d$ is the unknown parameter of interest, whereas the variables $\{\varepsilon_i\}_{i=1}^n$ correspond to additive noise. In robust regression, we obtain an estimate of the parameter vector μ^* by computing

$$\min_{\mu \in \mathbb{R}^d} \underbrace{\Big\{ \frac{1}{n} \sum_{i=1}^{n} \Psi\big(y_i - \langle z_i, \, \mu \rangle\big) \Big\}}_{=:\, f(\mu)} \tag{22}$$

where Ψ is a known loss function with some robustness properties. One popular example of the loss function Ψ is Tukey's bi-weight function, which is given by

$$\Psi(t) = \begin{cases} 1 - (1 - (t/\lambda)^2)^3 & \text{if } |t| \leq \lambda \\ 1 & \text{otherwise} \end{cases}, \tag{23}$$

where $\lambda > 0$ is a tuning parameter. Note that Ψ is a smooth function, whence the function f in the objective (22) is also smooth, implying that Algorithm 1 is suitable for the problem.

With this set-up, applying Theorem 1, Theorem 4 and Corollary 3, we obtain the following guarantee:

Corollary 4 *Given a random initialization, any bounded sequence* $\{\mu^k\}_{k \geq 0}$ *obtained by applying Algorithm 1 to the objective* (22) *has the following properties:*

(a) Almost surely with respect to the random initialization, the sequence $\{\mu^k\}_{k \geq 0}$ *converges to a point* $\bar{\mu}$ *such that* $\nabla f(\bar{\mu}) = 0$ *and* $\nabla^2 f(\bar{\mu}) \succeq 0$.

(b) There is a universal constant c_1 *such that*

$$\mathrm{Avg}\Big(\|\nabla f(\mu^k)\|_2\Big) \leq \frac{c_1}{k} \qquad \textit{for all } k = 1, 2, \ldots.$$

We provide the proof in Appendix F.1.

4.3. Smooth function minimization with sparsity constraints

Moving beyond the robust regression problem, we now discuss another interesting problem of minimizing a smooth function subject to sparsity penalty. Consider the following optimization problem

$$\min_{\substack{x \in \mathbb{R}^d \\ \|x\|_0 \leq s}} \quad g(x), \tag{24}$$

where g is a smooth function, the ℓ_0-"norm" $\|x\|_0$ counts the number of non-zero entries in the vector x, and $s \in \{1, \ldots, d\}$ is a sparsity parameter. The constraint set $\big\{x \in \mathbb{R}^d \mid \|x\|_0 \leq s\big\}$ is non-convex, and consequently, the optimization problem (24) is non-convex. However, the constraint set can be expressed as the level set of a certain DC function (Gotoh et al., 2017). In particular, let $|x|_{(d)} \geq |x|_{(d-1)} \geq \cdots \geq |x|_{(1)}$ denote the values of $x \in \mathbb{R}^d$ re-ordered in terms of their absolute magnitudes. In terms of this notation, we have $\|x\|_1 \geq \sum_{i=d-s+1}^{d} |x|_{(i)}$ for all $x \in \mathbb{R}^d$, with equality holding if and only if x is s–sparse. This fact ensures that

$$\Big\{x \in \mathbb{R}^d : \|x\|_0 \leq s\Big\} = \big\{x \in \mathbb{R}^d : \|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)} \leq 0\big\}. \tag{25}$$

Since both of the functions $x \mapsto \|x\|_1$ and $x \mapsto \sum_{i=d-s+1}^{d} |x|_{(i)}$ are convex (Boyd and Vandenberghe, 2004), this level set formulation is a DC constraint. Now using the representation (25), we can rewrite problem (24) as $\min_{x\in\mathbb{R}^d} g(x)$ such that $\|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)} \leq 0$. For our experiments, it is more convenient to solve the penalized analogue of the last problem, given by

$$\min_{x\in\mathbb{R}^d} \; \big\{g(x) + \lambda\Big(\|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)}\Big)\big\}, \tag{26}$$

where $\lambda > 0$ is a tuning parameter. The optimization problem (26) can be solved using Algorithm 2 with $g(x) = g(x)$, $\varphi(x) = \lambda\|x\|_1$ and $h(x) = \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$. For the non-smooth component $\varphi(x) = \lambda\|x\|_1$, there is a closed form expression of the proximal update in Algorithm 2, so that the method is especially efficient in this case.

4.3.1. Best subset selection

A special case of problem (26) arises from best subset selection in linear regression. Suppose that we observe a vector $y \in \mathbb{R}^n$ and a matrix $B \in \mathbb{R}^{n\times d}$ that are linked via the standard linear model $y = Bx^* + \varepsilon$. Here the vector $\varepsilon \in \mathbb{R}^n$ corresponds to additive noise, whereas $x^* \in \mathbb{R}^d$ is the unknown regression vector. We wish to estimate the unknown parameter vector x^* subject to a sparsity constraint, and we do so by solving the following optimization problem:

$$\min_{\substack{x\in\mathbb{R}^d \\ \|x\|_0\leq s}} \|y - Bx\|_2^2. \tag{27}$$

Here the non-negative integer s is a tuning parameter that controls maximum number of allowable non-zero entries in the vector x. Following the development leading to the formulation (26), let us consider instead the problem of minimizing the function

$$f(x) := \|y - Bx\|_2^2 + \lambda\Big(\|x\|_1 - \sum_{i=d-s+1}^{d} |x|_{(i)}\Big)\big\}. \tag{28}$$

Note that the function f can be decomposed as a difference of two convex functions as follows:

$$f(x) = \underbrace{\|y - Bx\|_2^2 + \lambda\|x\|_1}_{\text{convex}} - \lambda \underbrace{\sum_{i=d-s+1}^{d} |x|_{(i)}}_{\text{convex}}. \tag{29}$$

Consequently, problem (28) is a DC optimization problem; hence, it is amenable to standard DC optimization techniques like CCCP. We can also apply Algorithm 2 on problem (28) with $g(x) = \|y - Bx\|_2^2$, $\varphi(x) = \lambda\|x\|_1$ and $h(x) = \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$.

4.3.2. Comparison of Algorithm 2 and CCCP

Let us compare the performance of our Algorithm 2 (prox-type method) with the popular convex-concave procedure (CCCP) for minimizing differences of convex functions. We apply both algorithms to the best subset selection problem (28).

Let us reiterate that problem (28) can be written as a difference of two convex functions, and one can apply CCCP update (10) to the decomposition (29). The inner convex optimization problem in update (10) is solved by proximal methods for minimizing the sum of a smooth convex function and a ℓ_1 regularizer. We also apply Algorithm 2 on problem (28) with $g(x) = \|y - Bx\|_2^2$, $h(x) = \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$ and $\varphi(x) = \lambda\|x\|_1$.

Synthetic data generation: We generated the rows of the $n \times d$ matrix B from a d-dimensional Gaussian distribution with zero mean and an equicovariance matrix Σ, where $\Sigma_{ii} = 1$ for all i, and $\Sigma_{ij} = 0.7$ for all $i \neq j$. The regression vector $x^* \in \mathbb{R}^d$ (true value) was chosen to be a binary vector with sparsity s ($s \ll d$). The location of the nonzero entries of the vector x^* was chosen uniformly without replacement form the set $\{1, \ldots, d\}$.

Performance measures: We use the following two criteria to compare the performance of the prox-type method and CCCP.

(a) *Total runtime:* Firstly, we compare the algorithms in terms of their total runtime. The runtime was measured in units of seconds.

(b) *Estimation error:* Secondly, we use average estimation error of the algorithms as a measure of performance. Let us recall that if $\bar{x} \in \mathbb{R}^d$ is the estimated value of the unknown regression vector x^*, then the average estimation error is defined as $\frac{\|\bar{x}-x^*\|_2}{\sqrt{p}\|\bar{x}\|_2}$. Note that the average estimation error used here is invariant under scaling.

Comparison results: Figure 2 shows the performances of the prox-type method and CCCP for synthetic data simulated as above, with problem parameters $(n, p) = (190, 300)$ and $(n, p) = (380, 600)$ and different choices of sparsity s.

For both the algorithms, the tolerance level η was set to $\eta = 10^{-8}$, whereas the maximum number of iterations was 1000. Figure 2 suggests that total runtime of the prox-type method is significantly smaller than the runtime of CCCP. Furthermore, the estimation error for the prox-type method is lower compared to CCCP, which possibly suggests that prox-type method is finding better local minima compared to CCCP for the non-convex optimization problem (28). In all our simulations we used same initializations for both the algorithms. The simulation results shown in Figure 2 are average over 100 replications, and we also provide the pointwise error bar in the plots.

4.3.3. Some theoretical guarantees

Interestingly, it turns out that when applied to problem (28), the convergence behavior of Algorithm 2 to a given stationary point $\bar{x}$ depends on the behavior of a certain convex program defined in terms of $\bar{x}$. More precisely, for any point $\bar{x} \in \mathbb{R}^d$ with $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)}$, consider the following convex relaxation of problem (28):

$$\mathcal{P}(\bar{x}) := \min_{x \in \mathbb{R}^d} \left\{ \|y - Bx\|_2^2 + \lambda\|x\|_1 - \lambda\langle \nabla h(\bar{x}),\, x - \bar{x}\rangle \right\}. \tag{30}$$

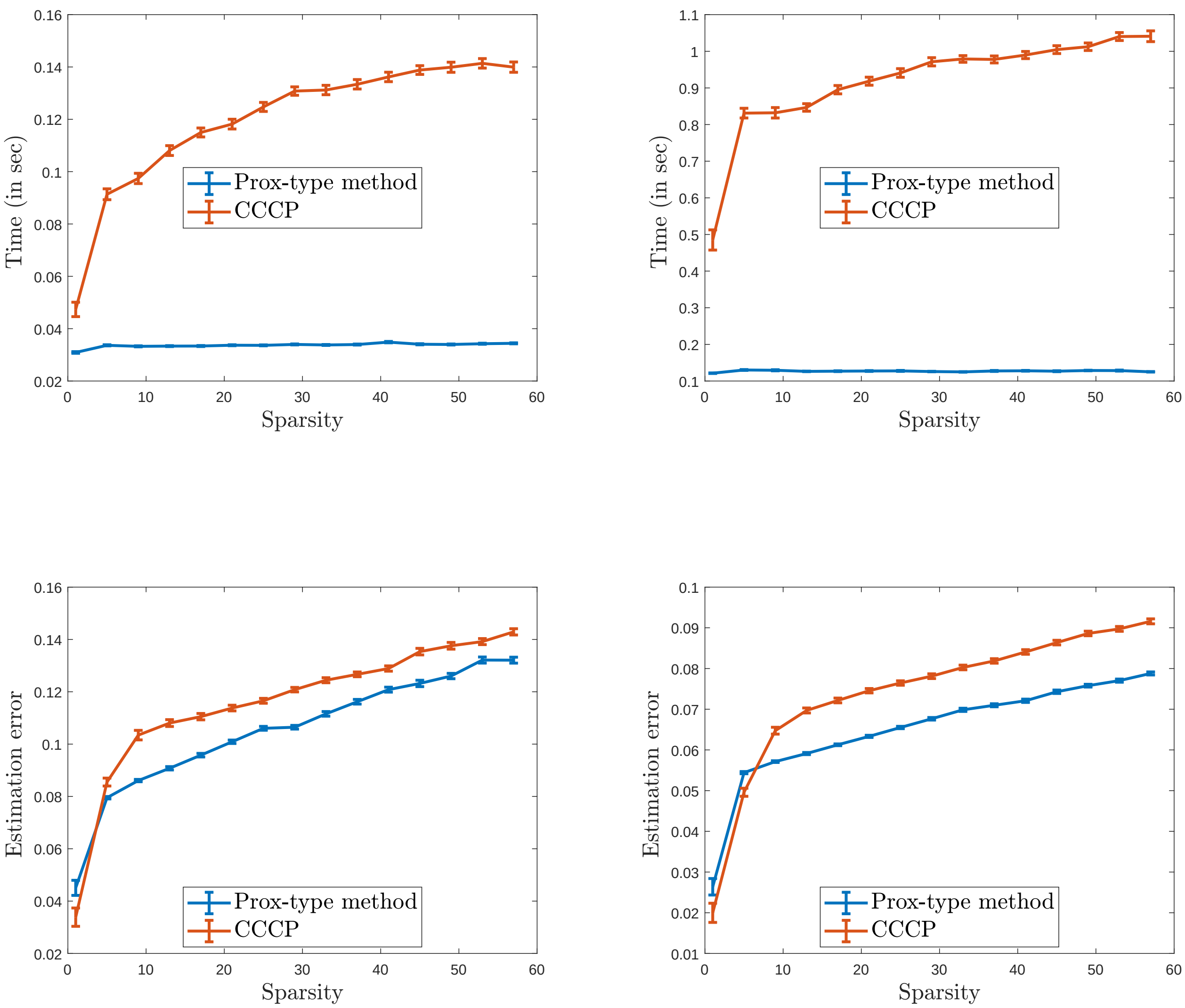

Figure 2. Performance of CCCP compared to that of Algorithm 2 on the best subset selection problem for synthetic data for different values of (n, p). The left columns correspond to $(n, p) = (190, 300)$, whereas the right columns correspond to $(n, p) = (380, 600)$. Plots in the first row compare the performance in terms of total runtime, those in the second row compare algorithms in terms of estimation error. We see that Algorithm 2 outperforms CCCP in terms of runtime. The performance of Algorithm 2 and CCCP in terms of estimation error are similar for low values of sparsity, whereas Algorithm 2 outperforms CCCP when sparsity is moderate to large. We initialized both the algorithms from the same starting point. Results shown above are averaged over 100 replications, and we also provide point-wise error bars in the plots.

Note that $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)}$ implies the differentiability of the function $h := \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$ which ensures that the above problem is well-defined.

Corollary 5 *Let $\{x^k\}_{k\geq 0}$ be any bounded sequence obtained by applying Algorithm 2 on problem* (28)*. Suppose there exists a limit point $\bar{x}$ of the sequence $\{x^k\}_{k\geq 0}$ satisfying $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)}$, and the convex problem* (30) *has unique solution. Then the sequence $\{x^k\}_{k\geq 0}$ converges to the point $\bar{x}$, and for all $k = 1, 2, \ldots$, we have*

$$\text{Avg}\left(\|\nabla f(x^k)\|_2\right) \leq \frac{c_1}{k}, \quad \text{and} \quad \|x^k - \bar{x}\|_2 \leq cq^k,$$

where $q \in (0, 1)$, and (c, c_1) are positive constants independent of k.

Comments on problem (30)**:** It can be shown that when the matrix B is of full rank, the objective function in problem (30) is strictly convex, and as a result, the problem (30) has unique solution. In the proof of Corollary 5, we show that the point $\bar{x}$ is always a minimizer of the convex problem (30), so that the uniqueness assumption implies that $\bar{x}$ is in fact the unique solution.

4.4. Mixture density estimation

As a final example, we consider the problem of estimating a two-component mixture density, where each of the constituent densities belong to an exponential family. The density of an exponential family (with respect to a fixed base measure, typically counting or Lebesgue) takes the form

$$p(y; \eta) = g(y) \exp\left\{\langle \eta, T(y)\rangle - A(\eta)\right\}. \tag{31}$$

Here the function $T : \mathcal{Y} \to \mathbb{R}^d$ is a vector of sufficient statistics, whereas the log-partition function

$$A(\eta) := \log\Big(\int_{\mathcal{Y}} g(y) \exp\{\langle \eta, T(y)\rangle\} dy\Big)$$

serves to normalize the density. The parameter vector $\eta \in \mathbb{R}^d$ determines the choice of density within the family. See Table 1 for some examples of 1-dimensional exponential families of this type. It includes various familiar examples, such as the Gaussian, Poisson and Beta families.

In the problem of mixture density estimation, one is interested in densities of the form

$$\zeta(y; \underbrace{\pi, \eta_0, \eta_1}_{\theta}) = \pi\, p(y; \eta_0) + (1 - \pi) p(y; \eta_1), \tag{32}$$

where $\pi \in (0, 1)$ is an unknown mixing proportion, and (η_0, η_1) are the unknown parameters of the two underlying densities.

Given n i.i.d. samples $\{y_i\}_{i=1}^n$ drawn from a mixture density of the form (32), a standard goal is to estimate the unknown parameter vector $\theta := (\pi, \eta_0, \eta_1)$. One way to do so is by

Distribution Name	η	$A(\eta)$	Twice continuously differentiable and sub-analytic
Poisson (λ)	$\ln(\lambda)$	$\exp\eta$	✓
Geometric (p)	$\ln(p)$	$-\ln(1-\exp\eta)$	✓
Gaussian(μ,σ^2)	$\left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right)^\top$	$-\frac{\eta_1^2}{4\eta_2} - \frac{1}{2}\ln(-2\eta_2)$	✓
Exponential (λ)	$-\lambda$	$-\ln(-\eta)$	✓
Gamma (α,β)	$(\alpha-1,\beta)^\top$	$\ln\Gamma(\eta_1+1)-(\eta_1+1)\ln(\eta_2)$	✓
Weibull (λ,k^5)	$-\frac{1}{\lambda^k}$	$\ln(-\eta)-\ln(k)$	✓
Beta (α,β)	$(\alpha,\beta)^\top$	$\ln\Gamma(\eta_1)+\ln\Gamma(\eta_2)-\ln\Gamma(\eta_1+\eta_2)$	✓

Table 1. Table showing the natural parameter η and the log-partition function A for different densities of exponential family, which are twice continuously differentiable and sub-analytic. In Appendix F.3 we prove the log-partition functions A mentioned in the above table are sub-analytic.

computing the maximum likelihood estimate (MLE), obtained via minimizing the negative log-likelihood of parameter θ given by the data. Frequently, a regularized form of the MLE is used, say of the form

$$\min_{\theta} \underbrace{\left\{ -\sum_{i=1}^{n} \log\big(\zeta(y_i;\theta)\big) \right\}}_{g(\theta)} \qquad \text{such that } \eta_0,\eta_1 \in \mathbb{R}^d,\ \pi\in[0,1], \text{ and } \|\eta_0\|_2 \le R_0,\ \|\eta_1\|_2 \le R_1. \tag{33}$$

Here $R_0 > 0$ and $R_1 > 0$ are tuning parameters providing upper bound on the ℓ_2-norms of the parameters η_0 and η_1 respectively, often chosen by a data-dependent procedure (such as cross-validation).

By inspection, the objective function g in problem (33) is non-convex. By standard theory on exponential families, the function A is always infinitely differentiable on its domain, so that the objective function g is infinitely differentiable on the convex set

$$\mathcal{X} = \Big\{\theta = (\eta_0,\eta_1,\pi) \mid \eta_j \in \mathrm{dom}(A), \pi\in[0,1], \|\eta_j\|_2 \le R_j \quad \text{for } j=0,1\Big\}.$$

Consequently, we may apply Algorithm 2 with $g(\cdot) = -\sum_{i=1}^{n} \log\big(\zeta(\cdot;y_i)\big)$, $h \equiv 0$ and $\varphi(\cdot) = \mathbb{1}_{\mathcal{X}}(\cdot)$ and $f = g - h + \varphi$. Interestingly, the log-partition function A is sub-analytic for many exponential family densities (see Table 1), which ensures that the function g is also sub-analytic. In Appendix A.3, we show that continuous sub-analytic functions satisfy Assumption KL so that we can apply Theorem 5 to obtain the following:

Corollary 6 *Any sequence* $\{\theta^k\}_{k\ge0} = \big\{\eta_0^k,\eta_1^k,\pi^k\big\}_{k\ge0}$ *obtained by applying Algorithm 2 to problem* (33) *satisfies the following properties:*

(a) It converges to a first order stationary point.

(b) *For all* $k = 1, 2, \ldots$, *we have* $\text{Avg}\left(\|\nabla f(\theta^k)\|_2\right) \leq \frac{c_1}{k}$, *where* c_1 *is a universal constant independent of* k.

See Appendix F.3 for the proof of this corollary.

5. Discussion

In this paper, we analyzed the behavior of three gradient-based algorithms—namely gradient descent, a proximal method, and an algorithm of the Frank-Wolfe type—for finding critical points of a class of non-convex non-smooth optimization problems. For each of the three algorithms, we provided non-asymptotic bounds on the rate of convergence to a first-order stationary point. We showed that our algorithm can escape strict saddle point for a class of non-smooth functions, thereby generalizing existing results for smooth functions. As a consequence of our theory, we obtained a simplification of the popular CCCP algorithm, and the simplified algorithm retains all the convergence properties of CCCP. Finally, we showed that for a large subclass of functions, which include continuous sub-analytic functions as a special case, we can have a significant improvement in the rate of convergence.

Our work leaves open a number of questions for future research. For instance, it would be interesting to characterize the class of DC-based functions mentioned in problem (2) when the convex function h is non-differentiable. Indeed, we then obtain a larger non-class of non-differentiable functions, and we suspect that Theorem 6 can be suitably generalized. Finally, we suspect that the proof techniques used here can be leveraged in order to establish sharper results for other forms of non-convex optimization problems.

Acknowledgments

This work was partially supported by the Office of Naval Research Grant DOD ONR-N00014 and National Science Foundation Grant NSF-DMS-1612948.

Appendix A. Technical background

In this appendix, we collect some technical background on subdifferentials and sub-analytic functions.

A.1. Fréchet and limiting subdifferential

We first recall the definitions and some useful properties of sub-differentials, which will be useful in subsequent sections.

Definition 1 *Let* $f : \mathbb{R}^d \mapsto \mathbb{R}$ *be a lower semicontinuous function. For any* $x \in \text{dom}(f)$, *the Fréchet subgradient of the function* f *at point* x *is defined as*

$$\widehat{\partial} f(x) = \left\{ u \;\Big|\; \liminf_{y \neq x, y \to x} \frac{f(y) - f(x) - \langle u, y - x \rangle}{\|y - x\|_2} \geq 0 \right\}.$$

Definition 2 *Let $f : \mathbb{R}^d \mapsto \mathbb{R}$ be a lower semi-continuous function. For any $x \in \mathrm{dom}(f)$, the limiting subdifferential of the function f at point x is defined as*

$$\partial_L f(x) = \Big\{ u \;\Big|\; \exists\;\, x^k \to x, u^k \to u \textit{ with } f(x^k) \to f(x) \textit{ and } u^k \in \widehat{\partial} f(x^k) \textit{ as } k \to \infty \Big\}.$$

Properties: The following properties of Fréchet and limiting sub-differential are provided in Chapter 8 of Rockafellar and Wets (2009).

(a) For any proper convex function h, we have $\partial_L h(x) = \widehat{\partial} h(x)$ for all $x \in \mathrm{dom}(h)$, and both quantities agree with the usual subgradient of the convex function h.

(b) If a function g is smooth in a neighborhood of a point x, then $\partial_L f(x) = \nabla f(x)$.

(c) Consider a function f of the form $f = g + \varphi$, where the function g is smooth in a neighborhood of a point x, and the function φ is proper convex and finite at the point x. Then the limiting sub-differential of the function f at the point x is given by $\partial_L f(x) = \nabla g(x) + \partial \varphi(x)$.

(d) *(Graph continuity:)* Consider a sequence $\big\{ \big(x^k, u^k\big) \big\}_{k \geq 1}$ in $\mathrm{graph}(\partial_L f)$ such that the sequence $\{(x^k, u^k, f(x^k)\}_{k \geq 0}$ converges to a point $(x, u, f(x))$. Then $(x, u) \in \mathrm{graph}(\partial_L f)$. Recall that $\mathrm{graph}(\partial_L f) := \big\{ (x, u) \in \mathbb{R}^d \times \mathbb{R} \mid u \in \partial_L f(x) \big\}$.

.

A.2. Sub-analytic functions satisfy KL-assumption

In this appendix, we show that continuous sub-analytic functions satisfy the KL-inequality. We also provide examples of functions which are sub-analytic.

Comments on limiting sub-differential: In order to facilitate our discussion, we mention some simple facts on limiting subdifferential of a function f, where f is of the form $f = g - h$ (Theorems 1 and 4) or $f = g + \varphi - h$ (Theorems 2 and 5). The following properties are direct consequences of properties of the limiting subdifferential mentioned in Appendix A.1.

- Suppose that the difference function $f = g - h$ satisfies parts (a) and (b) of Assumption GR. Then we have

$$\partial_L(-f)(x) = \partial h(x) - \nabla g(x), \quad \text{and moreover}$$
$$\|\nabla f(x)\|_2 := \|\nabla g(x) - \partial h(x)\|_2 = \|\partial_L(-f)(x)\|_2$$

- Suppose that the function $f = g + \varphi - h$, where the function h is locally smooth, and the function f satisfies Assumption PR part (b). Then $\partial_L f(x) = \nabla g(x) - \nabla h(x) + \partial \varphi(x)$. Consequently, we have that $\|\nabla f(x)\|_2 = \|\partial_L f(x)\|_2$.

We prove that continuous sub-analytic functions satisfy Assumption KL by exploiting results due to Bolte et al. (2007). Let us introduce some notation used in this paper. We use $m_f(x)$ to denote the ℓ_2 distance of the set $\partial_L f(x)$ from zero; concretely, $m_f(x) := \mathrm{dist}_{\|\cdot\|_2}\big(0, \partial_L f(x)\big)$. In Theorem 3.1 (for critical points of the function f) and Remark 3.2 (for non-critical points of the function f), Bolte et al. (2007) proved the following fact about sub-analytic functions.

Lemma 1 (Bolte et al. (2007)): *Let $f : \mathbb{R}^d \mapsto \mathbb{R} \cup \{+\infty\}$ be a sub-analytic function with closed domain, and assume that $f|_{\mathrm{dom}(f)}$ is continuous. Then for any $a \in \mathrm{dom}(f)$, there exists an exponent $\theta \in [0, 1)$ such that, the function $\frac{|f - f(a)|^\theta}{m_f}$ is bounded above in a neighborhood of a.*

Using Lemma 1, we now argue that sub-analytic functions, under the conditions of Theorem 4 or Theorem 5, satisfy Assumption KL.

Lemma 2 *Any sub-analytic function f satisfying Assumption GR also satisfies Assumption KL.*

Proof First, note that the function f is continuous by Assumption GR; suppose f is sub-analytic, then from properties of sub-analytic functions, we have that the function $-f$ is also sub-analytic. Furthermore, the function $-f$ is continuous in the closed domain $\mathcal{C}$ — which by Lemma 1 guarantees that, for any $a \in \mathcal{C}$, there exists $\theta \in [0, 1)$ such that the ratio $\frac{|-f-(-f(a))|^\theta}{m_{(-f)}}$ is bounded above in a neighborhood of the point a. Since $|-f-(-f(a))| = |f - f(a)|$, proving satisfiability of Assumption KL reduces to showing that $m_{(-f)}(x)$ is upper bounded by $\|\nabla f(x)\|_2$. To this end, note that from the discussion about limiting subdifferential in the paragraph above Lemma 1, we have

$$\|\nabla f(x)\|_2 = \|\partial_L(-f)(x)\|_2 \overset{(i)}{\geq} m_{(-f)}(x), \tag{34}$$

where step (i) follows from the definition of $m_{(-f)}(x)$. Putting together the pieces, we conclude that any sub-analytic function f which satisfies Assumption GR, also satisfies Assumption KL. ■

Lemma 3 *Suppose that, in addition to the conditions on the functions (g, h, φ) from Theorem 2, the function $f := g - h + \varphi$ is continuous and sub-analytic in its domain $\mathrm{dom}(f)$, and the domain $\mathrm{dom}(f)$ is closed. Then the function f satisfies Assumption KL.*

Proof Since the function $f|_{\mathrm{dom}(f)}$ is continuous and sub-analytic by assumption, from Lemma 1, we have that for any $a \in \mathrm{dom}(f)$ there exists a $\theta \in [0, 1)$ such that, the ratio $\frac{|f-f(a)|^\theta}{m_f}$ is bounded above in a neighborhood of the point a. In order to justify satisfiability of Assumption KL, it suffices to prove that $m_f(x)$ is upper bounded by $\|\nabla f(x)\|_2$. To this end, note that the function h is locally smooth by assumptions of Theorem 2 part (b). Hence, from the discussion about limiting subdifferential in the paragraph above Lemma 1, we have

$$\|\nabla f(x)\|_2 = \|\partial_L f(x)\|_2 \overset{(i)}{\geq} m_f(x), \tag{35}$$

where step (i) follows from the definition of $m_f(x)$. Putting together the pieces, guarantees that the function f satisfies Assumption KL. ■

A.3. Instances of sub-analytic functions

In Appendix A.2, we proved that continuous sub-analytic functions satisfy Assumption KL, and in those cases,—by Theorems 4 and 5—we have a faster rate of convergence of Algorithms 1 and 2. In this appendix, we provide examples of functions which are sub-analytic. We start by providing definitions of sub-analytic functions following the definition of (Bolte et al., 2007).

A subset $S \subset \mathbb{R}^d$ is called *semi-analytic* if each point of $\mathbb{R}^d$ admits a neighborhood V such that the set $S \cap V$ has the form

$$S \cap V = \cup_{i=1}^{p} \cap_{j=1}^{q} \big\{x \in V \mid h_{ij} = 0, g_{ij} > 0\big\},$$

where the functions $h_{ij}, g_{ij} : V \mapsto \mathbb{R}$ are real-analytic.

A set S is called *sub-analytic*, if each point of $\mathbb{R}^d$ admits a neighborhood V such that

$$S \cap V = \big\{x \in \mathbb{R}^d : \big(x, y\big) \in B\big\},$$

where B is a bounded semi-analytic subset of $\mathbb{R}^d \times \mathbb{R}^m$ for some $m \geq 1$. A function f is called sub-analytic if the graph of f, defined by $\text{graph}(f) := \big\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : f(x) = y\big\}$, is sub-analytic.

The class of sub-analytic functions is quite large. In order to motivate the reader, we provide few examples here. The following results can be found in Bolte et al. (2014) and Chapter 6 in the book by Facchinei and Pang (2007).

(a) Any real-valued polynomial or analytic function is sub-analytic.

(b) Any real-valued semi-algebraic or semi-analytic function is sub-analytic.

(c) Indicator function of a semi-algebraic set is sub-analytic.

(d) Sub-analytic functions are closed under finite linear combinations, and the product of two sub-analytic functions is sub-analytic.

(e) Point-wise maximum and minimum of a finite collection of sub-analytic functions are sub-analytic.

(f) *Composition rule:* If g_1 and g_2 are two sub-analytic functions with the function g_1 being continuous, then the composition function $g_2 \circ g_1$ is sub-analytic. In fact, the class of continuous sub-analytic functions are *closed under algebraic operations.*

Appendix B. Proofs related to Algorithm 1

In this appendix, we collect the proofs of various results related to the gradient-based Algorithm 1, including Theorem 1, Corollaries 1 and 3, and Proposition 1.

B.1. Proof of Theorem 1

Our proof of this theorem, as well as subsequent ones, depends on the following descent lemma:

Lemma 4 *Under the conditions of Theorem 1, we have*

$$x^k \in \mathrm{int}(\mathcal{C}) \quad \textit{and} \quad f(x^{k+1}) \le f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2 \quad \textit{for all } k = 0, 1, 2, \ldots. \tag{36}$$

See Appendix B.1.1 for the proof of this lemma.

We now prove Theorem 1 using Lemma 4.

Convergence of function values: We first prove that the function value sequence $\{f(x^k)\}_{k\ge 0}$ is convergent. Since $f^* := \min_{x\in\mathcal{C}} f(x)$ is finite by assumption, and $x^k \in \mathrm{int}(\mathcal{C})$ for all $k \ge 0$ by Lemma 4, the sequence $\{f(x^k)\}_{k\ge 0}$ is bounded below. For any non-stationary x^k, inequality (36) also ensures that $f(x^k) > f(x^{k+1})$; hence, there must exist some scalar $\bar{f}$ such that $\lim_{k\to\infty} f(x^k) = \bar{f}$.

Stationarity of limit points: Next, we establish that any limit point of the sequence $\{x^k\}_{k\ge 0}$ must be stationary. Consider a subsequence $\{x^{k_j}\}_{j\ge 0}$ of $\{x^k\}_{k\ge 0}$ such that $x^{k_j} \to \bar{x}$, and let $\{u^{k_j}\}_{j\ge 0}$ be the associated sequence of subgradients. It suffices to exhibit a subgradient $\bar{u} \in \partial h(\bar{x})$ such that $\nabla g(\bar{x}) - \bar{u} = 0$. Since the sequence $\{x^{k_j}\}_{j\ge 0}$ converges to $\bar{x}$, we must have

$$\|\nabla f(x^{k_j})\|_2 = \|\nabla g(x^{k_j}) - u^{k_j}\|_2 \to 0.$$

The function g is continuously differentiable by assumption, and we have $\nabla g(x^{k_j}) \to \nabla g(\bar{x})$. Combining these we find that $u^{k_j} \to \nabla g(\bar{x})$. Furthermore, by continuity of the function g, we have $g(x^{k_j}) \to g(\bar{x})$. Putting together the pieces we have established above that $\big(x^{k_j}, u^{k_j}, g(x^{k_j})\big) \to \big(\bar{x}, \bar{u}, g(\bar{x})\big)$, where $\bar{u} := \nabla g(\bar{x})$. Consequently, the graph continuity of limiting-sub-differentials (see Appendix A.1) guarantees that $\bar{u} = \nabla g(\bar{x}) \in \partial h(\bar{x})$. Overall, we conclude that $\nabla f(\bar{x}) := \nabla g(x) - \bar{u} = 0$, so that $\bar{x}$ is a stationary point as claimed.

Establishing the bound (3): Finally, we prove the claimed bound (3) on the averaged squared gradient. Recalling that $f^* := \min_{x\in\mathcal{C}} f(x)$ is finite, we have

$$\begin{aligned} f(x^0) - f^* \ge f(x^0) - f(x^{k+1}) &= \sum_{j=0}^{k} f(x^j) - f(x^{j+1}) \\ &\overset{(i)}{\ge} \frac{\alpha}{2}\sum_{j=0}^{k} \|\nabla f(x^k)\|_2^2 \\ &= \frac{\alpha(k+1)}{2}\mathrm{Avg}\left(\|\nabla f(x^k)\|_2^2\right), \end{aligned}$$

where step (i) follows from equation (36). Rearranging yields the claimed bound (3) on the averaged squared gradient.

B.1.1. Proof of Lemma 4

Recall that by assumption, the function g is continuously differentiable and M_g-smooth, and the function h is convex. As a consequence, for any vector $x^k \in \mathcal{C}$ and subgradient $u^k \in \partial h(x^k)$, we have

$$g(x) \le g(x^k) + \langle \nabla g(x^k),\, x - x^k \rangle + \frac{M_g}{2}\|x - x^k\|_2^2 \tag{37a}$$

$$h(x) \ge h(x^k) + \langle u^k,\, x - x^k \rangle. \tag{37b}$$

Combining inequalities (37a) and (37b) yield

$$f(x) = g(x) - h(x) \le f(x^k) + \langle \nabla g(x^k) - u^k,\, x - x^k \rangle + \frac{M_g}{2}\|x - x^k\|_2^2. \tag{38}$$

Substituting $x = x^{k+1} := x^k - \alpha\big(\nabla g(x^k) - u^k\big)$ in equation (38) and simplifying yields

$$f(x^k) - f(x^{k+1}) \ge \big(\frac{1}{\alpha} - \frac{M_g}{2}\big)\|x^{k+1} - x^k\|_2^2 = \alpha\big(1 - \frac{\alpha M_g}{2}\big)\|\nabla g(x^k) - u^k\|_2^2$$
$$\stackrel{(i)}{\ge} \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2,$$

where inequality (i) follows from the upper bound $\alpha \le \frac{1}{M_g}$. This proves the second part of the stated lemma. As for the claim that the sequence remains in the interior of the set $\mathcal{C}$, note that $f(x^{k+1}) \le f(x^k) \le f(x^0)$, which ensures that $x^{k+1} \in \mathcal{L}(f(x^0)) \subset \mathrm{int}(\mathcal{C})$, as claimed.

B.2. Proof of Corollary 1

The first part of the proof builds on a simple application of Theorem 1 and the definition of effective smoothness constant M_f^*. The second part of the proof utilizes a relation between the backtracking step size and the effective smoothness constant. For sake of completeness, we first describe the gradient descent backtracking algorithm.

Algorithm 4 Gradient descent with backtracking

1: Given an initial point $x^0 \in \mathrm{int}(\mathcal{C})$ and parameter $\beta \in (0, 1)$:
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Choose the smallest nonnegative integer i_k such that the step size $t^k := \beta^{i_k}$ satisfies:

$$f\big(x^k - t^k \nabla f(x^k)\big) \le f(x^k) - \frac{t^k}{2}\|\nabla f(x^k)\|^2. \tag{39}$$

4: Update $x^{k+1} = x^k - t^k \nabla f(x^k)$.
5: **end for**

Establishing the bound in (5a)**:** For any step size α in the interval $\big(0, \frac{1}{M_{f^*}}\big)$, the definition of the effective smoothness constant M_{f^*} ensures the following property. There exists a M_g-smooth function g and a convex-differentiable function h with $f = g - h$, and the scalar M_g satisfies $\alpha < \frac{1}{M_g} \leq \frac{1}{M_{f^*}}$. Since the function f is differentiable, applying Algorithm 1 on the function f with the decomposition $f = g - h$ is equivalent to applying gradient descent on f. Furthermore, the step size α satisfies the upper bound $\alpha \leq \frac{1}{M_g}$, and applying the bound (3) from Theorem 1 yields:

$$\operatorname{Avg}\Big(\|\nabla f(x^k)\|_2^2\Big) \leq \frac{2\big(f(x^0) - f^*\big)}{\alpha(k+1)}. \tag{40}$$

Establishing the backtracking bound (5b)**:** For any fraction $\beta \in (0,1)$, the definition of the effective smoothness constant M_{f^*} guarantees the following. There exists a M_g-smooth function g and a convex and differentiable function h with $f = g - h$, and the scalar M_g satisfies $\beta M_g \leq M_{f^*} \leq M_g$. Comparing the descent step (36) from Lemma 4 and step (39) in Algorithm 4, we conclude that the step size t^k satisfies the lower bound $t^k \geq \min\big\{1, \frac{\beta}{M_g}\big\} \geq \min\big\{1, \frac{\beta^2}{M_f^*}\big\}$. Applying the descent step (39) in Algorithm 4 repeatedly and then utilizing the last lower bound on step size t^k, we find that for all $k = 0, 1, 2\ldots$

$$f(x^0) - f(x^{k+1}) \geq \sum_{i=0}^{k} \frac{t^k}{2}\|\nabla f(x^k)\|^2 \geq \min\Big\{\frac{1}{2}, \frac{\beta^2}{2M_f^*}\Big\}\sum_{i=0}^{k}\|\nabla f(x^k)\|^2.$$

Rearranging the last inequality yields:

$$\begin{aligned}\operatorname{Avg}\Big(\|\nabla f(x^k)\|^2\Big) &\leq \frac{2\max\big\{1, \frac{M_f^*}{\beta^2}\big\}\big(f(x^0) - f(x^{k+1})\big)}{(k+1)} \\ &\overset{(i)}{\leq} \frac{2\max\big\{1, M_f^*\big\}\big(f(x^0) - f^*\big)}{\beta^2(k+1)},\end{aligned} \tag{41}$$

where step (i) follows since $\beta \in (0,1)$, along with the lower bound $f(x^{k+1}) \geq f^*$.

B.3. Proof of Corollary 3

Based on Theorem 4 of (Lee et al., 2016), it suffices to show that the gradient map $G(x) := x - \alpha\nabla f(x)$ is a diffeomorphism for any step size $\alpha \in \big(0, \frac{1}{M_g}\big)$. Recall that a map $G : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a diffeomorphism if the map G is a bijection, and both the maps G and G^{-1} are continuously differentiable.

Injectivity: We first prove that G is an injective map. Consider a pair of vectors x, y such that $G(x) = G(y)$; our aim is to prove that $x = y$. The condition $G(x) = G(y)$ is equivalent to $x - y = \alpha\big(\nabla f(x) - \nabla f(y)\big)$, and we have that

$$\begin{aligned}\|x - y\|_2^2 &= \alpha\langle x - y,\, \nabla f(x) - \nabla f(y)\rangle \\ &= \alpha\langle x - y,\, \nabla g(x) - \nabla g(y)\rangle - \alpha\langle x - y,\, \nabla h(x) - \nabla h(y)\rangle \\ &\overset{(i)}{\leq} \alpha M_g\|x - y\|_2^2 - \alpha\langle x - y,\, \nabla h(x) - \nabla h(y)\rangle \\ &\overset{(ii)}{\leq} \alpha M_g\|x - y\|_2^2.\end{aligned}$$

Here inequality (i) follows because the gradient ∇g is M_g-Lipschitz by assumption; inequality (ii) follows from the convexity of the function h, which implies the monotonicity of the gradient ∇h. Finally, since the step size $\alpha < \frac{1}{M_g}$ by assumption, the inequality $\|x - y\|_2^2 \leq \alpha M_g \|x - y\|_2^2$ can hold only when $x = y$.

Surjectivity: For any fixed vector $y \in \mathbb{R}^d$, consider the following problem

$$\arg\min_{x \in \mathbb{R}^d} \Big\{ \frac{1}{2}\|x - y\|_2^2 - \alpha g(x) + \alpha h(x) \Big\}. \tag{42}$$

Observe that for any step size $\alpha \in \big(0, \frac{1}{M_g}\big)$ and any fixed vector $y \in \mathbb{R}^d$, the map $x \mapsto \frac{1}{2}\|x - y\|_2^2 - \alpha g(x)$ is strongly convex, whence the map $x \mapsto \frac{1}{2}\|x - y\|_2^2 - \alpha g(x) + \alpha h(x)$ is also strongly convex. Consequently, the convex problem (42) has a unique minimizer, and we denote it by x_y. In order to prove surjectivity of the map G, it suffices to show the point x_y is mapped to the point y. Recalling the KKT conditions of the problem (42), we have that

$$y = x_y - \alpha \nabla f(x_y) = G(x_y),$$

which completes the proof of surjectivity of the map G.

Combining the injectivity and the surjectivity of the map G, we conclude that the inverse map G^{-1} exists. Next, let $DG(\cdot)$ denote the Jacobian of the map G, then $DG(x) = \mathrm{I} - \alpha \nabla^2 g(x) + \alpha \nabla^2 h(x)$. Since the function g is M_g-smooth, and the map G is continuously differentiable, standard application of the inverse-function theorem guarantees that for all step size $\alpha < \frac{1}{M_g}$, the inverse map G^{-1} is continuously differentiable. Putting together the pieces, we conclude that map G^{-1} exists, and both the maps (G, G^{-1}) are continuously differentiable. Overall, we have established that the map G is a diffeomorphism, as claimed.

B.4. Proof of Proposition 1

The CCCP update at step $(k+1)$ is given by $x^{k+1} = \arg\min_{x \in \mathcal{C}} q(x, x^k)$, where

$$q(x, x^k) := g(x) - h(x^k) - \langle \nabla h(x^k),\, x - x^k \rangle. \tag{43}$$

Observe that step $(k+1)$ of Algorithm 1 is equivalent to a gradient descent update with step size α on the map $x \mapsto q(x, x^k)$. Accordingly, if we define $y^{k+1} = x^k - \alpha \nabla q(x, x^k)$, then we have $q(y^{k+1}, x^k) \geq q(x^{k+1}, x^k)$; moreover

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\overset{(i)}{\geq} q(x^k, x^k) - q(x^{k+1}, x^k) \\ &\overset{(ii)}{\geq} q(x^k, x^k) - q(y^{k+1}, x^k) \\ &\overset{(iii)}{\geq} \frac{1}{2M_g}\|\nabla f(x^k)\|_2^2. \end{aligned} \tag{44}$$

Here inequality (i) follows from the equality $q(x^k, x^k) = f(x^k)$ combined with the lower bound $q(x, x^k) \geq f(x)$. Inequality (ii) follows since $q(y^{k+1}, x^k) \geq q(x^{k+1}, x^k)$, and inequality

(iii) follows from Lemma 4 with step size $\alpha = \frac{1}{M_g}$. Note that equation (44) guarantees that the function value sequence $\{f(x^k)\}_{k\geq 0}$ is decreasing. Since the function f is bounded below, we have that the sequence $\{f(x^k)\}_{k\geq 0}$ converges. In order to prove that all limit points of the sequence $\{x^k\}_{k\geq 0}$ are critical points, we follow the corresponding argument in proof of Theorem 1. This completes the proof of part (a) in Proposition 1.

Turning to part (b), unwrapping the recursive lower bound (44) and re-arranging yields inequality (11a). Finally, we turn to the proof of inequality (11b) under the additional strong convexity condition. Under this condition, the map $x \mapsto q(x, x^k)$ in equation (43) is μ-strongly convex, so that

$$f(x^k) - f(x^{k+1}) \geq q(x^k, x^k) - q(x^{k+1}, x^k) \overset{(i)}{\geq} \frac{\mu}{2}\|x^k - x^{k+1}\|_2^2, \tag{45}$$

where inequality (i) follows from the strong convexity of the map $x \mapsto q(x, x^k)$ and the fact that $\nabla q(x^{k+1}, x^k) = 0$. Using this equation repeatedly, we find that

$$\begin{aligned} f(x^0) - f^* \geq f(x^0) - f(x^{k+1}) &= \sum_{j=0}^{k} \left\{ f(x^j) - f(x^{j+1}) \right\} \\ &\geq \frac{\mu}{2} \sum_{j=0}^{k} \|x^j - x^{j+1}\|_2^2 \\ &= \frac{\mu(k+1)}{2} \text{Avg}\left(\|x^k - x^{k+1}\|_2^2 \right). \end{aligned}$$

Rearranging the last inequality yields the bound (11b). Finally, let us reiterate that bounds similar to (11b) are known in the literature; see the paper (Lanckriet and Sriperumbudur, 2009) for example. We provide the proof of bound (11b) for completeness.

Appendix C. Proof of Theorem 2

This proof shares some important steps with Theorem 1, but it requires a more refined argument due to the presence of a non-smooth and non-continuous function φ. We start by stating an auxiliary lemma that underlies the proof of Theorem 2. In the proof, the subgradients of the convex functions h and φ at a point x^k are denoted by u^k and v^k, respectively.

Lemma 5 *Under the conditions of Theorem 2, we have*

$$x^{k+1} = x^k - \alpha(\nabla g(x^k) + v^{k+1} - u^k), \qquad \textit{and} \tag{46a}$$

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2, \tag{46b}$$

valid for all $k = 0, 1, 2, \ldots$. Furthermore, for any convergent subsequence $\{x^{k_j}\}_{j\geq 0}$ of the sequence $\{x^k\}_{k\geq 0}$ with $x^{k_j} \to \bar{x}$, we have

$$\lim_{j\to\infty} \varphi(x^{k_j+1}) = \varphi(\bar{x}).$$

See Appendix C.1 for the proof of this lemma.

We now prove Theorem 2 using Lemma 5.

Convergence of function value: We first prove that the sequence of function values $\{f(x^k)\}_{k\geq 0}$ is convergent. Since $f^* := \min_{x\in\mathbb{R}^d} f(x)$ is finite by assumption, the sequence $\{f(x^k)\}_{k\geq 0}$ is bounded below. If $x^k = x^{k+1}$ for some k, the convergence of the sequence $\left\{f(x^k)\right\}_{k\geq 0}$ is trivial. Hence, we may assume without loss of generality that $x^k \neq x^{k+1}$ for all $k = 0, 1, 2, ...$ In that case, inequality (46b) ensures that $f(x^k) > f(x^{k+1})$, and consequently, there must exist some scalar $\bar{f}$ such that $\lim_{k\to\infty} f(x^k) = \bar{f}$.

Stationarity of limit points: Next, we establish that any limit point of the sequence $\left\{x^k\right\}_{k\geq 0}$ must be stationary. Consider a subsequence $\left\{x^{k_j}\right\}_{j\geq 0}$ such that $x^{k_j} \to \bar{x}$. Let $\left\{v^{k_j}\right\}_{j\geq 0}$ and $\left\{u^{k_j}\right\}_{j\geq 0}$ be the associated sequence of subgradients. It suffices to exhibit subgradients $\bar{v} \in \partial\varphi(\bar{x})$ and $\bar{u} \in \partial h(\bar{x})$ such that, $\nabla g(\bar{x}) + \bar{v} - \bar{u} = 0$.

Step 1: *Existence of subgradient* $\bar{u}$: Since the sequence $\left\{x^{k_j}\right\}_{j\geq 0}$ is convergent, we may assume that the sequence $\left\{x^{k_j}\right\}_{j\geq 0}$ is bounded, and it lies in a compact set S. The function h is convex continuous, and we have that $h(x^{k_j}) \to h(\bar{x})$, and the subgradient sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ is bounded; see example 9.14 in the book (Rockafellar and Wets, 2009). Passing to a subsequence if necessary, we may assume that the sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ converges to $\bar{u}$. Putting together these pieces, we conclude that $(x^{k_j}, u^{k_j}, h(x^{k_j})) \to (\bar{x}, \bar{u}, h(\bar{x}))$ as $j \to \infty$; consequently, the graph continuity of limiting sub-differentials guarantees that $\bar{u} \in \partial h(\bar{x})$ (see Appendix A.1 for graph continuity).

Step 2: *Existence of subgradient* $\bar{v}$: In order to complete the proof, it suffices to show that the vector $\bar{v} := -\nabla g(\bar{x}) + \bar{u}$ belongs to the subgradient set $\partial\varphi(\bar{x})$. Since the norm of successive difference $\|x^{k_j} - x^{k_j+1}\|_2$ converges to zero, Lemma 5 yields $\|\nabla g(x^{k_j}) + v^{k_j+1} - u^{k_j}\|_2 \to 0$, and $x^{k_j+1} \to \bar{x}$. Furthermore, continuity of the gradient ∇g yields $\nabla g(x^{k_j}) \to \nabla g(\bar{x})$, and step 1 above guarantees $u^{k_j} \to \bar{u}$. Combining these two facts with $\|\nabla g(x^{k_j}) + v^{k_j+1} - u^{k_j}\|_2 \to 0$, we obtain $v^{k_j+1} \to \bar{v} := -\nabla g(\bar{x}) + \bar{u}$, and by Lemma 5, we have $\varphi(x^{k_j+1}) \to \varphi(\bar{x})$. Putting together the pieces, we conclude that $(x^{k_j+1}, v^{k_j+1}, \varphi(x^{k_j+1})) \to (\bar{x}, \bar{v}, \varphi(\bar{x}))$. Consequently, the graph continuity of limiting subdifferentials guarantees that $\bar{v} \in \partial\varphi(\bar{x})$ (see Appendix A.1 for graph continuity).

Finally, the subgradients $\bar{u} \in \partial h(\bar{x})$ and $\bar{v} \in \partial\varphi(\bar{x})$ obtained from steps 1 and and 2 respectively satisfy the relation $\nabla g(\bar{x}) + \bar{v} - \bar{u} = 0$, which establishes the claimed stationarity of $\bar{x}$.

Establishing the bound (13a)**:** Next, we establish the claimed bound (13a) on the averaged squared successive difference. Recalling that $f^* := \min_{x \in \mathbb{R}^d} f(x)$ is finite, we have

$$\begin{aligned} f(x^0) - f^* \geq f(x^0) - f(x^{k+1}) &= \sum_{j=0}^{k} f(x^j) - f(x^{j+1}) \\ &\overset{(i)}{\geq} \frac{1}{2\alpha} \sum_{j=0}^{k} \|x^j - x^{j+1}\|_2^2 \\ &= \frac{(k+1)}{2\alpha} \text{Avg}\Big(\|x^k - x^{k+1}\|_2^2\Big), \end{aligned} \tag{47}$$

where step (i) follows from equation (46b). Rearranging the last inequality yields the claimed bound (13a) on the averaged squared successive difference.

Establishing the bound (13b)**:** In order to establish the bound (13b) on the averaged squared gradient, we start by establishing the following upper bound on the gradient-norm $\|\nabla f(x^{k+1})\|_2$:

$$\|\nabla f(x^{k+1})\|_2 \leq \big(M_g + M_h + \frac{1}{\alpha}\big)\|x^k - x^{k+1}\|_2. \tag{48}$$

Recall that the function h is M_h smooth by assumption, and we have

$$\begin{aligned} \|\nabla g(x^{k+1}) - \nabla h(x^{k+1}) + v^{k+1}\|_2 &\overset{(i)}{=} \|\nabla g(x^{k+1}) - \nabla h(x^{k+1}) + \big(\nabla h(x^k) - \nabla g(x^k) + \frac{1}{\alpha}\big(x^k - x^{k+1}\big)\big)\|_2 \\ &\overset{(ii)}{\leq} \|\nabla g(x^k) - \nabla g(x^{k+1})\|_2 + \|\nabla h(x^k) - \nabla h(x^{k+1})\|_2 + \frac{1}{\alpha}\|x^k - x^{k+1}\|_2 \\ &\overset{(iii)}{\leq} \big(M_g + M_h + \frac{1}{\alpha}\big)\|x^k - x^{k+1}\|_2. \end{aligned}$$

Here step (i) follows from the update equation of x^{k+1} in Lemma 5 and from differentiability of the function g; step (ii) follows from triangle inequality, and step (iii) follows from the smoothness of the functions g and h. Putting together the bounds (48) and (47), we obtain the desired bound (13b).

C.1. Proof of Lemma 5

Here we prove the claims of Lemma 5.

Establishing update equation (46a)**:** Recalling the convex majorant defined in equation (38), we define a convex majorant $q(\cdot, x^k)$ of the function f as follows:

$$q(x, x^k) = g(x^k) - h(x^k) + \langle \nabla g(x^k) - u^k, x - x^k \rangle + \frac{1}{2\alpha}\|x - x^k\|_2^2 + \varphi(x), \tag{49}$$

where subgradient $u^k \in \partial h(x^k)$, and the step size α satisfies $0 < \alpha \leq \frac{1}{M_g}$. Observe that minimizer of the convex function $x \mapsto q(x, x^k)$ over $x \in \mathbb{R}^d$ is same as $\text{prox}_{1/\alpha}^{\varphi}\big(x^k - \alpha(\nabla g(x^k) - u^k)\big)$, which implies that x^{k+1} is a minimizer of the convex function $x \mapsto q(x, x^k)$

over $x \in \mathbb{R}^d$. Consequently, the optimality condition of x^{k+1} guarantees that there exists subgradient $v^{k+1} \in \partial g(x^{k+1})$ satisfying the following equation:

$$\nabla g(x^k) - u^k + v^{k+1} + \frac{1}{\alpha}\big(x^{k+1} - x^k\big) = 0. \tag{50}$$

Rewriting the above equation yields the update equation (46a).

Establishing the descent step (46b)**:** Note that

$$\begin{aligned} f(x^k) - q(x^{k+1}, x^k) &\overset{(i)}{\geq} g(x^k) - h(x^k) + \varphi(x^{k+1}) + \langle v^{k+1},\, x^k - x^{k+1}\rangle - q(x^{k+1}, x^k) \\ &\overset{(ii)}{\geq} \langle \nabla g(x^k) - u^k + v^{k+1},\, x^k - x^{k+1}\rangle - \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\ &\overset{(iii)}{\geq} \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2. \end{aligned} \tag{51}$$

Here step (i) follows from the convexity of the function φ; step (ii) follows by substituting $q(x^{k+1}, x^k)$ from equation (49). In step (iii), we use the relation $\nabla g(x^k) - u^k + v^{k+1} = \frac{1}{\alpha}\big(x^k - x^{k+1}\big)$, which follows from equation (50). Finally, recall that the function $x \mapsto q(x, x^k)$ is a majorant for the function f, and we deduce that

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\geq f(x^k) - q(x^{k+1}, x^k) \\ &\geq \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2. \end{aligned} \tag{52}$$

Limit of the sequence $\big\{\varphi(x^{k_j+1})\big\}_{j\geq 0}$**:** Consider any convergent subsequence $\big\{x^{k_j}\big\}_{j\geq 0}$ of the sequence $\big\{x^k\big\}_{k\geq 0}$ with $x^{k_j} \to \bar{x}$. Recall that $f^* = \inf_{x\in\mathbb{R}^d} f(x)$ is finite by assumption; combining this with step (46b) in Lemma 5, we have that $\|x^k - x^{k+1}\|_2 \to 0$, and that $x^{k_j+1} \to \bar{x}$. The function φ is lower semi-continuous, and we have

$$\liminf_{j\to\infty}\ \varphi(x^{k_j+1}) \geq \varphi(\bar{x}). \tag{53}$$

Since we already proved x^{k_j+1} is a minimizer of the convex function $x \mapsto q(x, x^{k_j})$, we have $q(x^{k_j+1}, x^{k_j}) \leq q(\bar{x}, x^{k_j})$. Unwrapping the last inequality and taking lim sup yields

$$\begin{aligned} \limsup_{j\to\infty}\ \varphi(x^{k_j+1}) &\overset{(i)}{\leq} \varphi(\bar{x}) + \limsup_{j\to\infty}\ \Big(\langle \bar{x} - x^{k_j},\, \nabla g(x^{k_j}) - u^{k_j}\rangle + \frac{1}{2\alpha}\|x^{k_j} - \bar{x}\|_2^2\Big) \\ &\overset{(ii)}{=} \varphi(\bar{x}). \end{aligned} \tag{54}$$

Here step (i) holds since $\|x^{k_j} - x^{k_j+1}\|_2 \to 0$, and the sequence $\big\{\nabla g(x^{k_j})\big\} - u^{k_j}\big\}_{j\geq 0}$ is bounded—which we prove shortly; step (ii) above follows from $x^{k_j} \to \bar{x}$ and boundedness of the sequence $\big\{\nabla g(x^{k_j}) - u^{k_j}\big\}_{j\geq 0}$. Combining equations (53) and (54) we obtain the claimed result.

Boundedness of the sequence $\left\{\nabla g(x^{k_j}) - u^{k_j}\right\}_{j\geq 0}$: In order to prove the boundedness of the sequence $\left\{\nabla g(x^{k_j}) - u^{k_j}\right\}_{j\geq 0}$, it suffices to show that the gradient sequence $\left\{\nabla g(x^{k_j})\right\}_{j\geq 0}$ and the sub-gradient sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ are bounded. Recall that $x^{k_j} \to \bar{x}$, and we have that the sequence $\{x^{k_j}\}_{j\geq 0}$ is bounded. Consequently, from the smoothness of the function g, we find that the sequence $\left\{\nabla g(x^{k_j})\right\}_{j\geq 0}$ is bounded. Finally, note that the function h is convex continuous, and we already argued that the sequence $\{x^{k_j}\}_{j\geq 0}$ is bounded. Combining this with example 9.14 in the book (Rockafellar and Wets, 2009), we conclude that the subgradient sequence $\left\{u^{k_j}\right\}_{j\geq 0}$ bounded.

Appendix D. Proofs related to Algorithm 3

In this appendix, we provide the proof of Theorem 3, which applies to the Frank-Wolfe based method (Algorithm 3). We also provide an upper bound on the generalized curvature constant $\mathcal{C}_f$, which is stated in Lemma 6.

D.1. Proof of Theorem 3

Let $x^\gamma := x^k + \gamma d^k$, where the difference d^k is defined as $d^k := s^k - x^k$, and the vector s^k is the Frank-Wolfe direction defined in Algorithm 3. Unpacking the definition (18) of the generalized curvature constant $\mathcal{C}_f$, we find that for any scalar $\gamma \in (0, 1)$ and subgradient $u^k \in \partial h(x^k)$, we have the following:

$$\begin{aligned} f(x^\gamma) &\leq f(x^k) + \gamma\langle \nabla g(x^k) - u^k,\, d^k\rangle + \frac{\gamma^2}{2}\mathcal{C}_f \\ &\overset{(i)}{\leq} f(x^k) - \gamma g^k + \frac{\gamma^2}{2} C_0. \end{aligned} \tag{55}$$

Here inequality (i) is obtained by substituting $g^k = \langle d^k,\, u^k - \nabla g(x^k)\rangle$ and using $C_0 \geq \mathcal{C}_f$. Substituting $\gamma = \gamma^k := \min\left\{\frac{g^k}{C_0}, 1\right\}$ in equation (55) yields

$$f(x^{k+1}) \leq f(x^k) - \min\left\{\frac{(g^k)^2}{2C_0}, g^k - \frac{C_0}{2}\mathbb{1}_{\left\{g^k > C_0\right\}}\right\}, \tag{56}$$

where $x^{k+1} = x^k + \gamma^k d^k$. Let $\bar{g}^k := \min_{0\leq j\leq k} g^j$ denote the minimum FW gap up to iteration k, then repeated application of equation (56) yields

$$\begin{aligned} f(x^0) - f(x^{k+1}) &\geq \sum_{j=0}^{k} \min\left\{\frac{(g^j)^2}{2C_0}, g^j - \frac{C_0}{2}\mathbb{1}_{\left\{g^j > C_0\right\}}\right\} \\ &\geq (k+1)\min\left\{\frac{(\bar{g}^k)^2}{2C_0}, \bar{g}^k - \frac{C_0}{2}\mathbb{1}_{\left\{\bar{g}^k > C_0\right\}}\right\}. \end{aligned} \tag{57}$$

Rewriting the last equation yields the following upper bound

$$\min\left\{\frac{(\bar{g}^k)^2}{2C_0}, \bar{g}^k - \frac{C_0}{2}\mathbb{1}_{\left\{\bar{g}^k > C_0\right\}}\right\} \overset{(i)}{\leq} \frac{f(x^0) - f^*}{k+1},$$

where step (i) follows from the lower bound $f(x^{k+1}) \geq f^* := \min_{x \in \mathcal{C}} \; f(x)$. Considering the cases where $\bar{g}^k \leq C_0$ and $\bar{g}^k > C_0$ separately, it can be shown following (Lacoste-Julien, 2016) that

$$\bar{g}^k \leq \begin{cases} \frac{2(f(x^0)-f^*)}{\sqrt{k+1}} & \text{for } \; k+1 \leq \frac{2(f(x^0)-f^*)}{C_0} \\ \\ \sqrt{\frac{2C(f(x^0)-f^*)}{k+1}} & \text{otherwise} \; . \end{cases}$$

Finally, note that $\sqrt{2C_0(f(x^0)-f^*)} \leq \max\{2(f(x^0)-f^*), C_0\}$ and we conclude that

$$\bar{g}^k \leq \frac{\max\left\{2\big(f(x^0)-f^*\big), C_0\right\}}{\sqrt{k+1}}.$$

D.2. Upper bound on generalized curvature constant

In this section, we provide an upper bound on the generalized curvature constant $\mathcal{C}_f$, where the function f is a difference of a differentiable function g and a continuous function h. For better readability, we use $\mathcal{C}_{g-h}$ instead of $\mathcal{C}_f$ in the following lemma.

Lemma 6 *Suppose that the function g is continuously differentiable and function h is convex, then we have $\mathcal{C}_{g-h} \leq \mathcal{C}_g$. Furthermore, if the function g is M_g-smooth, and the function h is a μ strongly convex function with $0 \leq \mu < M$, then*

$$\mathcal{C}_{g-h} \leq \big(M - \mu\big) \times \big(\operatorname{diam}_{\|\cdot\|_2}(\mathcal{C})\big)^2, \tag{58}$$

where $\operatorname{diam}_{\|\cdot\|_2}$ *denote the diameter of the set $\mathcal{C}$, measured in ℓ_2 norm.*

Comments: The first upper bound on $\mathcal{C}_{g-h}$ in Lemma 6 posits that the curvature constant of the difference function $g-h$ is upper bounded by curvature constant of the function g, whenever the second function h is convex. Let us try to understand an implication of this result through an example. One of the well-known upper bound of curvature constant for M_g-smooth function g is $M_g \times \big(\operatorname{diam}_{\|\cdot\|_2}(\mathcal{C})\big)^2$; see the paper by (Jaggi, 2013). Now consider continuously differentiable functions g and h such that the function g is M_g-smooth and the function h is non-smooth and convex. It can be verified that the difference function $g-h$ is *not* smooth in this case; consequently, the earlier bound on curvature constant $\mathcal{C}_{g-h}$ is ∞, whereas Lemma 6 ensures that

$$\mathcal{C}_{g-h} \leq \mathcal{C}_g \leq M_g \times \big(\operatorname{diam}_{\|\cdot\|_2}(\mathcal{C})\big)^2.$$

Proof of the upper bound $\mathcal{C}_{g-h} \leq \mathcal{C}_g$: Unwrapping the definition of $\mathcal{C}_{g-h}$, we have

$$\begin{aligned}
\mathcal{C}_{g-h} &= \sup_{\substack{x,y \in c_\gamma \\ u \in \partial h(x)}} \frac{2}{\gamma^2}\big[f(y) - f(x) - \langle y-x, \, \nabla g(x) - u\rangle\big] \\
&= \sup_{\substack{x,y \in c_\gamma \\ u \in \partial h(x)}} \frac{2}{\gamma^2}\big[g(y) - g(x) - \langle y-x, \, \nabla g(x)\rangle - \Delta_h(y,x,u)\big] \qquad (59) \\
&\overset{(i)}{\leq} \underbrace{\sup_{x,y \in c_\gamma} \frac{2}{\gamma^2}\big[f(y) - f(x) - \langle y-x, \, \nabla g(x)\rangle\big]}_{\mathcal{C}_g},
\end{aligned}$$

where $\Delta_h(y, x, u) := h(y) - h(x) - \langle y - x,\, u\rangle$. Here inequality (i) follows by noting that, for any pair of points $x, y \in \mathcal{C}$, and for any convex function h with $u \in \partial h(x)$, we have $\Delta_h(y, x, u) \geq 0$.

Proof of upper bound (58)**:** Suppose in addition, the function g is M_g-smooth, and the function h is μ-strongly convex with $\mu \geq 0$. Then we have $\Delta_h(y, x, u) \geq \frac{\mu}{2}\|x - y\|_2^2$, and equation (59) yields

$$\begin{aligned}\mathcal{C}_{g-h} &\leq \sup_{x,y \in c_\gamma} \frac{2}{\gamma^2}\big[g(y) - g(x) - \langle y - x,\, \nabla g(x)\rangle - \frac{\mu}{2}\|x - y\|_2^2\big] \\ &\overset{(i)}{\leq} \sup_{x,y \in c_\gamma} \frac{2}{\gamma^2}\left[\frac{M_g - \mu}{2}\|x - y\|_2^2\right],\end{aligned}$$

where step (i) follows since the function g is M_g-smooth. Substituting $y - x = \gamma s$ with $s \in \mathcal{C}$, we obtain the claimed upper bound

$$\mathcal{C}_{g-h} \leq (M_g - \mu) \times \big(\operatorname{diam}_{\|\cdot\|_2}(\mathcal{C})\big)^2.$$

Appendix E. Proofs of faster rates under Assumption KL

In this appendix, we prove our results on improved convergence rates for functions which satisfy Assumption KL—as stated in Theorems 4 and 5. We begin by stating an auxiliary lemma that underlies the proofs of Theorems 4 and 5.

Lemma 7 *Under assumptions of either Theorem 4 or Theorem 5, there exists constants* $\theta \in [0, 1)$*,* $C > 0$ *and positive integer* k_1 *such that for all* $k \geq k_1$*, we have*

$$|f(x^k) - \bar{f}|^\theta \leq C\|\nabla f(x^k)\|_2,$$

where $f(x^k) \downarrow \bar{f}$*. Furthermore, if* $x^k \to \bar{x}$*, then the parameters* (θ, C)*, obtained from KL-inequality of the function* f *at the point* $\bar{x}$*, satisfy the above inequality.*

See Appendix E.3 for the proof of this lemma.

E.1. Proof of Theorem 4

Now we prove Theorem 4 using Lemma 7.

Convergence of the sequence $\{x^k\}_{k\geq 0}$ **:** We demonstrate the convergence of the sequence $\{x^k\}_{k\geq 0}$ by proving that the sequence has finite length property; more precisely, we show that $\sum_{k=0}^{\infty}\|x^k - x^{k+1}\|_2 < \infty$. First, note that for any scalar $0 \leq \theta < 1$, the function $t \mapsto t^{1-\gamma\theta}$ is concave for $0 < \gamma < \frac{1}{\theta}$; consequently, for iteration $k \geq k_1$ we have

$$\begin{aligned}\big(f(x^k) - \bar{f}\big)^{1-\gamma\theta} - \big(f(x^{k+1}) - \bar{f}\big)^{1-\gamma\theta} &\geq \big(1 - \gamma\theta\big)\big(f(x^k) - \bar{f}\big)^{-\gamma\theta}\big(f(x^k) - f(x^{k+1})\big) \\ &\overset{(i)}{\geq} \big(1 - \gamma\theta\big)\big(|f(x^k) - \bar{f}|\big)^{-\gamma\theta} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\ &\overset{(ii)}{\geq} \frac{(1 - \gamma\theta)}{C\|\nabla f(x^k)\|_2^\gamma} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\ &\overset{(iii)}{=} \frac{(1 - \gamma\theta)}{2C\alpha^{1-\gamma}}\|x^k - x^{k+1}\|_2^{2-\gamma}. \qquad (60)\end{aligned}$$

Here inequality (i) follows from the descent property in equation (36) and from the fact that $f(x^k) \downarrow \bar{f}$. Inequality (ii) follows from Lemma 7, and equality (iii) follows from the relation $x^k - x^{k+1} = \alpha\big(\nabla g(x^k) - u^k\big) = \alpha \nabla f(x^k)$. Substituting $\gamma = 1$ and summing both side of inequality (60) from index $k = k_1$ to $k = \infty$, we obtain

$$\begin{aligned}\big(f\big(x^{k_1}\big) - \bar{f}\big)^{1-\theta} &= \sum_{k=k_1}^{\infty} \big(f\big(x^k\big) - \bar{f}\big)^{1-\theta} - \big(f\big(x^{k+1}\big) - \bar{f}\big)^{1-\theta} \\ &\geq \sum_{k=k_1}^{\infty} \frac{(1-\theta)}{2C} \|x^k - x^{k+1}\|_2,\end{aligned}$$

which proves the finite length property of the sequence $\{x^k\}_{k\geq 0}$. Consequently, we are guaranteed to have a vector $\bar{x}$ such that $x^k \to \bar{x}$ as $k \to \infty$.

Rate of convergence of Avg $(\|\nabla f(x^k)\|_2)$: Rewriting equation (60), we have the following:

$$\begin{aligned}C_\gamma := & \sum_{\ell=0}^{k_1} \frac{(1-\gamma\theta)}{2C\alpha^{1-\gamma}} \|x^\ell - x^{\ell+1}\|_2^{2-\gamma} + \big(f\big(x^{k_1}\big) - \bar{f}\big)^{(1-\gamma\theta)} \\ \stackrel{(i)}{\geq} & \sum_{\ell=0}^{k-1} \frac{(1-\gamma\theta)}{2C\alpha^{1-\gamma}} \|x^\ell - x^{\ell+1}\|_2^{2-\gamma} \\ = & \frac{k(1-\gamma\theta)}{2C\alpha^{1-\gamma}} \text{Avg}\left(\|x^k - x^{k+1}\|_2^{2-\gamma}\right),\end{aligned} \tag{61}$$

where step (i) above follows from equation (60), and $\text{Avg}\left(\|x^k - x^{k+1}\|_2^{2-\gamma}\right) := \frac{1}{k}\sum_{\ell=0}^{k-1} \|x^\ell - x^{\ell+1}\|_2^{2-\gamma}$ denote the running arithmetic average. Since $0 \leq \theta < 1$, we can take $\gamma = 1$ in equation (61), and we obtain the following rate:

$$\text{Avg}\left(\|\nabla f(x^k)\|_2\right) = \frac{1}{\alpha}\text{Avg}\left(\|x^k - x^{k+1}\|_2\right) \leq \frac{c_1}{k},$$

where $c_1 = \frac{2CC_\gamma}{\alpha(1-\theta)}$. Finally, note that the last equality holds trivially for iteration $k \leq k_1$ with the given choice of the constant c_1.

Rate of convergence of GAvg $(\|\nabla f(x^k)\|_2)$: Since we proved that the sequence $\{x^k\}_{k\geq 0}$ is convergent to the point $\bar{x}$, we have that the parameter θ in Lemma 7 can be taken to be the KL-exponent of the function f at point $\bar{x}$. Suppose $\frac{1}{2} \leq \theta < \frac{r}{2r-1}$, then substituting $\gamma = \frac{2r-1}{r}$ in equation (61) yields,

$$\begin{aligned}\text{GAvg}\left(\|\nabla f(x^k)\|_2\right) &= \frac{1}{\alpha}\text{GAvg}\left(\|x^k - x^{k+1}\|_2\right) \\ &\stackrel{(i)}{\leq} \frac{1}{\alpha}\left\{\text{Avg}\left(\|x^k - x^{k+1}\|_2^{\frac{1}{r}}\right)\right\}^r \\ &\stackrel{(ii)}{\leq} \frac{c_2}{k^r},\end{aligned}$$

where $c_2 = \frac{1}{\alpha}\big(\frac{2CC_\gamma\alpha^{1-\gamma\theta}}{1-\gamma\theta}\big)^r$ with $\gamma = \frac{2r-1}{r}$, and $\mathrm{GAvg}\left(\|x^k - x^{k+1}\|_2^{2-\gamma}\right) := \prod_{\ell=0}^{k-1}\big(\|x^\ell - x^{\ell+1}\|_2\big)^{\frac{1}{k}}$, the geometric average of the sequence $\big\{\|x^\ell - x^{\ell+1}\|_2\big\}_{l=0}^{k-1}$. Here step (i) above follows from arithmetic-geometric mean (AM/GM) inequality; step (ii) follows from the bound in equation (61) and from the fact that $\gamma = \frac{2r-1}{r}$. Finally, note that the last equality holds trivially for iteration $k \le k_1$ with the given choice of constant c_2.

E.2. Proof of Theorem 5

The proof of Theorem 5 builds on the techniques used in the proof of Theorem 4 but requires additional technical care due to the presence of possibly non-continuous function φ.

Convergence of the sequence $\{x^k\}_{k\ge 0}$: The proof of Theorem 5 has two steps. First, we prove a descent condition similar to equation (60). We then leverage this descent condition and weighted AM-GM inequality to obtain the desired result.

Step 1: Following the proof of Theorem 4, we prove the convergence of the sequence $\{x^k\}_{k\ge 0}$ by showing that the sequence $\{x^k\}_{k\ge 0}$ has finite length property. First, note that for scalars $0 \le \theta < 1$ and $0 < \gamma < \frac{1}{\theta}$, the function $t \mapsto t^{1-\gamma\theta}$ is concave. Consequently, for iteration $k \ge k_1$, from Lemma 7 we have

$$\begin{aligned}\big(f(x^k) - \bar{f}\big)^{1-\gamma\theta} - \big(f(x^{k+1}) - \bar{f}\big)^{1-\gamma\theta} &\ge \big(1-\gamma\theta\big)\big(f(x^k) - \bar{f}\big)^{-\gamma\theta}\big(f(x^k) - f(x^{k+1})\big) \\ &\overset{(i)}{\ge} \big(1-\gamma\theta\big)\big(|f(x^k) - \bar{f}|\big)^{-\gamma\theta} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\ &\overset{(ii)}{\ge} \frac{(1-\gamma\theta)}{C\|\nabla f(x^k)\|_2^\gamma} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2. \end{aligned} \tag{62}$$

Here step (i) follows from the descent property in equation (52) and from the fact that $f(x^k) \downarrow \bar{f}$; step (ii) follows from Lemma 7. The function h is locally smooth by assumption; as a result, we have that the difference function $g - h$ is locally smooth. We also assumed that the sequence $\{x^k\}_{k\ge 0}$ is bounded (lies in a compact set S); consequently, we may assume that the difference function $g-h$ is smooth in the compact set S with a smoothness parameter M_{g-h}(say). Borrowing the argument of Theorem 2 part(b), it follows that:

$$\|\nabla g(x^k) - \nabla h(x^k) + v^k\|_2 \le \big(M_{g-h} + \frac{1}{\alpha}\big)\|x^k - x^{k-1}\|_2. \tag{63}$$

Combining the last inequality with inequality (62) yields the following descent property

$$\big(f(x^k) - \bar{f}\big)^{1-\gamma\theta} - \big(f(x^{k+1}) - \bar{f}\big)^{1-\gamma\theta} \ge \frac{(1-\gamma\theta)}{2\alpha C\big(M_{g-h} + \frac{1}{\alpha}\big)^\gamma} \times \frac{\|x^k - x^{k+1}\|_2^2}{\|x^k - x^{k-1}\|_2^\gamma}. \tag{64}$$

Step 2: We now leverage the descent condition obtained from step 1 to prove finite length property of the sequence $\{x^k\}_{k\ge 0}$. In order to facilitate further discussion, we use Δ_γ^k to denote the following:

$$\Delta_\gamma^k := C_3\Big(\big(f(x^k) - \bar{f}\big)^{1-\gamma\theta} - \big(f(x^{k+1}) - \bar{f}\big)^{1-\gamma\theta}\Big),$$

where the constant $C_3 := \frac{2\alpha C\left(M_{g-h}+\frac{1}{\alpha}\right)^{\gamma}}{(1-\gamma\theta)}$. With this notation, we can rewrite the equation (64) as

$$\Delta_\gamma^k \|x^{k-1} - x^k\|_2^\gamma \geq \|x^k - x^{k+1}\|_2^2. \tag{65}$$

Combining equation (65) with the weighted AM-GM inequality, we obtain

$$\begin{aligned}
\big(1 + \frac{\gamma}{2-\gamma}\big) \times \sum_{j=k_1+1}^{k} \|x^j - x^{j+1}\|_2^{2-\gamma} &\overset{(i)}{\leq} \big(1 + \frac{\gamma}{2-\gamma}\big) \times \sum_{k=k_1+1}^{k} \Big(\sqrt{\Delta_\gamma^j \, \|x^{j-1} - x^j\|_2^\gamma}\Big)^{\frac{2-\gamma}{2}} \\
&\overset{(ii)}{\leq} \sum_{j=k_1+1}^{k} \big(\Delta_\gamma^j + \frac{\gamma}{2-\gamma}\|x^{j-1} - x^j\|_2^{2-\gamma}\big) \\
&\overset{(iii)}{\leq} C_3\big(f(x^{k_1}) - \bar{f}\big)^{1-\gamma\theta} + \sum_{j=k_1+1}^{k} \frac{\gamma}{2-\gamma}\|x^{j-1} - x^j\|_2^{2-\gamma}.
\end{aligned} \tag{66}$$

Here step (i) follows from equation (65), and step (ii) is implied by applying weighted AM-GM inequality as follows:

$$\frac{\Delta_\gamma^j + \frac{\gamma}{2-\gamma}\|x^{j-1} - x^j\|_2^{2-\gamma}}{1 + \frac{\gamma}{2-\gamma}} \geq \Big(\Delta_\gamma^j \|x^{j-1} - x^j\|_2^\gamma\Big)^{\frac{1}{1+\frac{\gamma}{2-\gamma}}}.$$

Step (iii) in equation (66) follows from the following observation

$$\begin{aligned}
\sum_{j=k_1}^{k} \Delta_\gamma^j &= C_3 \sum_{j=k_1}^{k} \big(f(x^j) - \bar{f}\big)^{1-\gamma\theta} - \big(f(x^{j+1}) - \bar{f}\big)^{1-\gamma\theta} \\
&\leq C_3\big(f(x^{k_1}) - \bar{f}\big)^{1-\gamma\theta}.
\end{aligned}$$

Rewriting inequality (66), we have for all $k \geq k_1 + 2$

$$\begin{aligned}
\sum_{j=k_1+1}^{k-1} \|x^j - x^{j+1}\|_2^{2-\gamma} &\leq C_3\big(f(x^{k_1}) - \bar{f}\big)^{1-\gamma\theta} + \frac{\gamma}{2-\gamma}\|x^{k_1} - x^{k_1+1}\|_2^{2-\gamma} - \big(1 + \frac{\gamma}{2-\gamma}\big)\|x^k - x^{k+1}\|_2^{2-\gamma} \\
&\leq C_3\big(f(x^{k_1}) - \bar{f}\big)^{1-\gamma\theta} + \frac{\gamma}{2-\gamma}\|x^{k_1} - x^{k_1+1}\|_2^{2-\gamma} < \infty.
\end{aligned} \tag{67}$$

Finally, by substituting $\gamma = 1$ and letting $k \to \infty$ in the last equation, we deduce the finite length property of the sequence $\{x^k\}_{k\geq 0}$.

Rate of convergence of Avg $\big(\|\nabla f(x^k)\|_2\big)$ and GAvg $\big(\|\nabla f(x^k)\|_2\big)$: The proof of this part follows from the corresponding proof in Theorem 4 and using the inequality (67) and upper bound (63).

E.3. Proof of Lemma 7

Since the sequence $\{x^k\}_{k\geq 0}$ is bounded by assumption, without loss of generality, we may assume that the set of limit points of the sequence $\{x^k\}_{k\geq 0}$ — which we denote by $\bar{\mathcal{X}}$ — is a compact set. From Theorem 1 (respectively Theorem 2), we have that all the limit points of the sequence $\{x^k\}_{k\geq 0}$ are critical points of the function f; furthermore, since $f(x^k) \downarrow \bar{f}$, we also have that the function f is constant on the set of limit points $\bar{\mathcal{X}}$, and the function value on $\bar{\mathcal{X}}$ equals $\bar{f}$. Combining this with Assumption KL, we have for all $z \in \bar{\mathcal{X}}$, there exists constants $\theta(z) \in [0, 1)$, $r_z > 0$ and $C(z) > 0$ such that, $| f(x) - \bar{f} |^{\theta(z)} \leq C(z) \times \|\nabla f(x)\|_2$ for all $x \in B(z, r_z)$. Now, consider the open cover $\{B(z, r_z) : z \in \bar{\mathcal{X}}\}$ of the set $\bar{\mathcal{X}}$. From compactness of the set $\bar{\mathcal{X}}$, we are guaranteed to have a finite subcover; more precisely, there exists $\{z_1, \dots z_p\} \subseteq \bar{\mathcal{X}}$ such that $\bar{\mathcal{X}} \subseteq \bigcup_{i=1}^{p} B(z_i, r_{z_i})$. Define constants $\theta := \max\{\theta(z_i) : 1 \leq i \leq p\}$, $C := \max\{C(z_i) : 1 \leq i \leq p\}$, and $r := \min\left\{\frac{r_{z_i}}{2} : 1 \leq i \leq p\right\}$. Utilizing the result $\|x^k - x^{k+1}\|_2 \to 0$ from Theorem 1 (respectively Theorem 2), one can show that, there exists positive integer k_1 such that for all $k \geq k_1$ we have $\|x^k - x^{k+1}\|_2 < \frac{r}{2}$, and $x^k \in \bigcup_{i=1}^{p} B(z_i, r_{z_i})$. Putting together these pieces, we conclude that for all $k \geq k_1$

$$x^k \in \bigcup_{i=1}^{p} B(z_i, r_{z_i}), \quad \text{and} \quad | f(x^k) - \bar{f} |^{\theta} \leq C \|\nabla f\|_2,$$

which proves the first part of claimed lemma. Now suppose the sequence $\{x^k\}_{k\geq 0}$ converges to a point $\bar{x}$, then we have that the set of limit points $\bar{\mathcal{X}} = \{\bar{x}\}$, is a singleton set. The rest of the proof is immediate by repeating the argument so far, with the additional information that $\bar{\mathcal{X}} = \{\bar{x}\}$.

Appendix F. Proofs of Corollaries

In this appendix, we collect the proofs of Corollaries 4, 5 and 6 from Section 4.

F.1. Proof of Corollary 4

First, note that in order to apply Theorem 1 and Theorem 4 to Corollary 4, it is enough to show that the function $\mu \mapsto f(\mu)$ is M_f-smooth (in this example, function $h \equiv 0$, and hence $f \equiv g$), and the function f satisfies Assumption KL. We verify that Assumption KL is satisfied by proving that the objective function f in problem (22) is continuous sub-analytic (see Appendix A.2). For proving sub-analyticity, we heavily use the properties mentioned in Appendix A.3. In the following proof, we assume without loss of generality that $\lambda = 1$.

The function f is continuous sub-analytic: First, we show that the function Ψ is sub-analytic. We begin by observing that Ψ is piecewise polynomial. Polynomials are analytic functions and intervals are semi-analytic sets. Since piecewise analytic functions with semi-analytic pieces are semi-analytic (hence sub-analytic), we conclude that the function Ψ is sub-analytic. Now, the function $\mu \mapsto y_i - \langle z_i, \mu\rangle$ is linear, and hence continuous sub-analytic. Furthermore, since continuous sub-analytic functions are closed under composition, we have that the function $\mu \mapsto \Psi(y_i - \langle z_i, \mu\rangle)$ is sub-analytic. Finally, note that sub-analytic functions are closed under linear combination, and we conclude that the function f is sub-analytic. The continuity of the function f is immediate by inspection.

The function f is smooth: Since the vectors $\{(z_i, y_i)\}_{i=1}^n$ are fixed, it suffices to prove that the function Ψ is smooth. A straightforward calculation shows that Ψ is continuously differentiable and smooth; in particular, it has a smoothness parameter 36 when $\lambda = 1$.

Putting together the pieces, we conclude that Theorem 1 and Theorem 4 are applicable for problem (22). Convergence of the sequence $\{\mu^k\}_{k\geq 0}$ to a point $\bar{\mu}$ and the convergence rate of gradient norms follows from Theorem 4, and the stationary condition $\nabla f(\bar{\mu}) = 0$ follows from Theorem 1.

Escaping strict saddle points: Note that the functions (g, h) are twice continuously differentiable, and the function g is smooth. Consequently, from Corollary 3, it follows that with random initializations, Algorithm 1 avoids strict saddle points almost surely.

F.2. Proof of Corollary 5

We begin by providing a high-level outline of the proof. First, note that from Theorem 2, we have the successive difference $\|x^k - x^{k+1}\|_2 \to 0$, and as a result, the set of limit point of the sequence $\{x^k\}_{k\geq 0}$—call it $\bar{\mathcal{X}}$—is a connected set (Ostrowski, 2016). We prove that the connected-set $\bar{\mathcal{X}}$ is singleton by showing that the set $\bar{\mathcal{X}}$ has an isolated point — this also proves that sequence $\{x^k\}_{k\geq 0}$ is convergent. Next, we show that the objective-function f, in the problem (28), satisfies Assumption KL with exponent $\theta = \frac{1}{2}$. Finally, we show that condition $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)} \geq 0$ implies that function $x \mapsto h(x) := \sum_{i=d-s+1}^{d} |x|_{(i)}$ is smooth in a neighborhood of point $\bar{x}$, and we use the proof techniques of Theorem 5 to establish the convergence rate of the gradient sequence. In order to obtain the rate of convergence of the sequence $\{x^k\}_{k\geq 0}$, we use ideas similar to those in the paper (Lee et al., 2016).

Convergence of the sequence $\{x^k\}_{k\geq 0}$: For notational convenience, let us use $g(x) := \|y - Bx\|_2^2$, $\varphi(x) := \lambda\|x\|_1$, and $h(x) := \lambda \sum_{i=d-s+1}^{d} |x|_{(i)}$. Since the point $\bar{x}$ satisfies the condition $|\bar{x}|_{(r)} > |\bar{x}|_{(r+1)} \geq 0$ by assumption, there must exist a neighborhood $B(\bar{x}, r)$ such that the function h is differentiable in the neighborhood $B(\bar{x}, r)$, and all points $x \in B(\bar{x}, r)$ satisfy $\text{sign}(x_{(i)}) = \text{sign}(\bar{x}_{(i)})$ for $1 \leq i \leq r$. We show that, in a neighborhood of the point $\bar{x}$, it is the only critical point, thereby proving that the point $\bar{x}$ is an isolated critical point. To this end consider the convex sub-problem mentioned in Corollary 5

$$\mathcal{P}(\bar{x}) := \min_{x\in\mathbb{R}^d} g(x) + \lambda\varphi(x) - \lambda\langle\nabla h(\bar{x}), x\rangle. \tag{68}$$

For any point x^* such that $x^* \in B(\bar{x}, r) \cap \bar{\mathcal{X}}$, from Theorem 2, we know that

$$\nabla g(\bar{x}) + \lambda\bar{u} - \lambda\nabla h(\bar{x}) = 0 \quad \text{and} \quad \nabla g(x^*) + \lambda u^* - \lambda\nabla h(x^*) = 0, \tag{69}$$

where subgradients $u^* \in \partial\varphi(x^*)$ and $\bar{u} \in \partial\varphi(\bar{x})$. Next, note that from the choice of neighborhood $B(\bar{x}, r)$, it follows that for all $x \in B(\bar{x}, r)$ we have $\nabla h(x) = \nabla h(\bar{x})$, and in particular, we deduce $\nabla h(x^*) = \nabla h(\bar{x})$. Combining this relation with equation (69) yields:

$$\nabla g(\bar{x}) + \lambda\bar{u} - \lambda\nabla h(\bar{x}) = 0 \quad \text{and} \quad \nabla g(x^*) + \lambda u^* - \lambda\nabla h(\bar{x}) = 0,$$

which implies both the points x^* and $\bar{x}$ are zero sub-gradient points of convex problem (68); this contradicts the assumption that problem (68) has an unique solution. Hence, we conclude that $x^* = \bar{x}$, and the point $\bar{x}$ is an isolated critical point of the sequence $\{x^k\}_{k\geq 0}$, and $\bar{\mathcal{X}}$. Putting together the pieces, we conclude that $x^k \to \bar{x}$.

Smoothness of function h in a neighborhood of $\bar{x}$: We already argued above that for all $x \in B(\bar{x}, r)$, the function h is differentiable and $\nabla h(x) = \nabla h(\bar{x})$. Consequently, we have that in the neighborhood $B(\bar{x}, r)$, the function h is smooth with a smoothness parameter $M_h = 0$.

The function f satisfies Assumption KL with exponent $\theta = \frac{1}{2}$: Recently, in the paper (Li and Pong, 2016) (Corollaries 5.1 and 5.2), the authors showed that if the functions $f_1, f_2, \ldots, f_T$ satisfy the KL-inequality with an exponent $\theta = \frac{1}{2}$, then the function $f := \min\{f_1, f_2, \ldots, f_T\}$ also satisfies KL-inequality with the exponent $\theta = \frac{1}{2}$. Interestingly, the function f can be represented as is minimum of finitely many functions as follows:

$$f(x) = \min_{a \in \mathcal{A}} \{\|y - Bx\|_2^2 + \lambda\|x\|_1 - \lambda a^\top x\}, \tag{70}$$

where $\mathcal{A} := \{a \in \{-1, 0, 1\}^d : \sum_{i=1}^d |a_i| = r\}$. Note that the set $\mathcal{A}$ has cardinality at most 3^d. It is known that functions of the form $x \mapsto \frac{1}{2}x^\top Ax + P(x) + b^\top x$ satisfy the KL-inequality with exponent $\theta = \frac{1}{2}$, where P is a proper closed polyhedral function, and A is a positive semi-definite matrix; see Corollaries 5.1 and 5.2 in the paper (Li and Pong, 2016). Putting together these two observations, we conclude that the function f satisfies KL-assumption with KL-exponent $\theta = \frac{1}{2}$.

Combining the pieces: Since we proved $x^k \to \bar{x}$, we have that for a suitable choice of k_1, the tail sequence $\{x^k\}_{k \geq k_1}$ lies in the neighborhood $B(\bar{x}, r)$. Now, the function f satisfies Assumption KL with exponent $\theta = \frac{1}{2}$, and the function h is smooth in the neighborhood $B(\bar{x}, r)$; hence, following the argument in proof of Theorem 5 part(b), we conclude that:

$$\text{Avg}\Big(\|\nabla f(x^k)\|_2\Big) \leq \frac{c_1}{k}.$$

Rate of convergence of sequence $\{x^k\}_{k\geq 0}$: The KL-exponent for the function f is $\theta = \frac{1}{2}$, and we may use $\gamma = 1$ in equation (67) which yields

$$\sum_{\ell=k_1+1}^{\infty} \|x^\ell - x^{\ell+1}\|_2 \leq \|x^{k_1} - x^{k_1+1}\|_2 + C_3\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}}, \tag{71}$$

for some constant C_3. From Lemma 7 and equation (48), we have

$$\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \leq C\|\nabla f(x^{k_1})\|_2 \leq C(M + M_h + \frac{1}{\alpha})\|x^{k_1} - x^{k_1-1}\|_2. \tag{72}$$

Combining equations (71) and (72) we have

$$\begin{aligned}\sum_{\ell=k_1}^{\infty} \|x^\ell - x^{\ell+1}\|_2 &\leq 2\|x^{k_1} - x^{k_1+1}\|_2 + C_3\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \\ &\overset{(i)}{\leq} 2\|x^{k_1} - x^{k_1+1}\|_2 + CC_3(M + M_h + \frac{1}{\alpha})\|x^{k_1} - x^{k_1-1}\|_2 \\ &\overset{(ii)}{\leq} \bar{C}\|x^{k_1} - x^{k_1-1}\|_2, \end{aligned}\tag{73}$$

where $\bar{C}$ is a constant depending on M, M_h, α, C_3 and C, and step (i) above follows from equation (72). We justify step (ii) shortly, but let us first derive the linear rate of convergence of the sequence $\{x^k\}_{k\geq 0}$ using the derivation in equation (73). Denote $e_k = \sum_{\ell=k}^{\infty} \|x^\ell - x^{\ell+1}\|_2$. Then equation (73) provides the following recursion

$$e_{k_1} \leq \bar{C}(e_{k_1-1} - e_{k_1}).$$

Simple inspection of proof of Theorem 5 and derivations so far ensure that we can derive the equations (71) and (72) for all $k \geq k_1$; this provides us a recursion relation as above with k_1 replaced by k. Furthermore, by choosing a larger value of the constant $\bar{C}$ if necessary, we may conclude that for all $k \geq 1$ we have

$$e_k \leq \bar{C}(e_{k-1} - e_k).$$

Rearranging the above inequality yields $e_k \leq \frac{\bar{C}}{\bar{C}+1} e_{k-1}$, which guarantees that the sequence $\{e_k\}_{k\geq 0}$ converges to zero at a linear rate. Finally, observe that $\|x^k - x^*\|_2 \leq \sum_{\ell=k}^{\infty} \|x^\ell - x^{\ell+1}\|_2 = e_k$, and the linear rate of convergence of the sequence $\{\|x^k - x^*\|_2\}_{k\geq 0}$ to zero follows.

Justification for step (ii) in equation (73)**:** Note that it suffices to show that the object $\|x^{k_1} - x^{k_1+1}\|_2$ is upper bounded by a constant multiple of $\|x^{k_1} - x^{k_1-1}\|_2$, where the constant depends only on M, M_h, α and C. Recalling the decent property proved in equation (52) we have:

$$\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \geq \big(f(x^{k_1}) - f(x^{k_1+1})\big)^{\frac{1}{2}} \geq \frac{1}{\sqrt{2\alpha}} \|x^{k_1} - x^{k_1+1}\|_2. \tag{74}$$

Combining equations (74) and (72) we obtain the following upper and lower bound of $\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}}$:

$$\frac{1}{\sqrt{2\alpha}} \|x^{k_1} - x^{k_1+1}\|_2 \leq \big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \leq C(M + M_h + 1/\alpha)\|x^{k_1} - x^{k_1-1}\|_2.$$

Rearranging the last equality proves the desired upper bound. Finally, we reiterate that the above justification also hold for any iterate k with $k \geq k_1$.

F.3. Proof of Corollary 6

The proof of this corollary is based on application of Theorems 2 and 5. We verify the assumptions of Theorems 2 and 5 with $g(\theta) = -\sum_{i=1}^{n} \log\big(\zeta(y_i; \theta)\big)$, $h \equiv 0$, $\varphi = \mathbb{1}_{\mathcal{X}}$ and function $f := g - \varphi + h$. Note that the domain $\text{dom}(f) = \mathcal{X}$ is compact, which guarantees that the iterate sequence $\{\theta^k\}_{k\geq 0}$ obtained from Algorithm 2 is bounded. The function $h \equiv 0$ is smooth. The log-partition function A is twice continuously differentiable by assumption, which guarantees that the function g is also twice continuously differentiable, whence smooth in the compact domain $\mathcal{X}$. Finally, we verify that the function f satisfies Assumption KL by proving that f is continuous sub-analytic in its domain $\mathcal{X}$, and the domain $\mathcal{X}$ is closed; see Lemma 3. Clearly, $\text{dom}(f) = \mathcal{X}$ is closed, and the function f is continuous in $\text{dom}(f)$. Finally, we show that the functions (g, φ) are sub-analytic, and invoking the property (d) of sub-analytic functions form Appendix A.3, we conclude that the function $f := g + \varphi$ is sub-analytic.

The function φ is sub-analytic: Here, we use a simple result by (Attouch et al., 2010), which states that the indicator function of a semi-algebraic set is a semi-algebraic function (hence a sub-analytic function). In order to show that the set $\mathcal{X}$ is semi-algebraic, we note the following representation of the set $\mathcal{X}$

$$\mathcal{X} = \Big\{ \sum_{i=1}^{d} \theta_i^2 > R_1^2 \Big\}^c \bigcap \Big\{ \sum_{i=d+1}^{2d} \theta_i^2 > R_2^2 \Big\}^c \bigcap \big\{ \theta_{2d+1} > 1 \big\}^c \bigcap \big\{ -\theta_{2d+1} > 0 \big\}^c. \tag{75}$$

Each of the four sets in representation (75) are semi-algebraic by definition, and semi-algebraic sets are closed under finite intersection and complements; see the book by (Coste, 2002). Putting together these two observations, we conclude that the set $\mathcal{X}$ is semi-algebraic, and that $\mathbb{1}_{\mathcal{X}}$ is a sub-analytic function.

The function g is sub-analytic: The log-partition function A is sub-analytic by assumption. For a fixed vector y, the map $\eta \mapsto \eta^\top T(y)$ is linear, and hence sub-analytic. Since sub-analytic functions are closed under a finite linear combination, we conclude that the map $\eta \mapsto \eta^\top T(y) - A(\eta)$ is sub-analytic. Continuous sub-analytic functions are closed under multiplication and composition; since the $\exp(\cdot)$ function is continuous sub-analytic, we have for every fixed vector y the following map

$$(\eta_0, \eta_1, p) \mapsto \zeta(y; \eta_0, \eta_1, p) := p \exp(\eta_0^\top T(y) - A(\eta_0)) + (1-p) \exp(\eta_1^\top T(y) - A(\eta_1))$$

is sub-analytic. Furthermore, the $\log(\cdot)$ function analytic on the interval $(0, \infty)$, and using the composition rule for continuous sub-analytic functions, we obtain that the map $\theta \mapsto \log(\zeta(y_i; \theta))$ is sub-analytic, where $\theta := (\eta_0, \eta_1, p)$. Finally, the target function g is a linear combination of sub-analytic functions $\log(\zeta(y_i; \theta))$, and we conclude that the map $\theta \mapsto g(\theta)$ is sub-analytic.

Combining the pieces: Putting together the pieces, we conclude that the function f is sub-analytic, with the function f being continuous in $\mathrm{dom}(f)$, whereas $\mathrm{dom}(f)$ is closed; furthermore, the functions g and h are smooth. This allows us to apply Theorem 2 and Theorem 5 and the corollary follows.

Sub-analyticity of the log-partition functions A in Table 1: The sub-analyticity of the log-partition function A mentioned in Table 1 follows from the following two observations. First, note that the functions $\exp, \ln$ and Γ are continuous and analytic (hence sub-analytic). Given two continuous sub-analytic functions g_1 and g_2, the composition function $g_2 \circ g_1$ is also continuous sub-analytic. Secondly, any linear combination of sub-analytic functions is also sub-analytic function. See Appendix A.3 for properties of sub-analytic functions.

Appendix G. Characterizing "smooth - convex" function class

In Theorem 1 and Theorem 2 we discussed a class of non-smooth non-convex functions, where a gradient or a prox-type algorithm provides satisfactory convergence to a critical point. One possible deficiency of the theory discussed so far is that, in Algorithm 1 (respectively Algorithm 2), we need to specify a decomposition of the objective function f as a difference of a smooth and a convex function (respectively, smooth + convex - convex). Consequently, it is natural to wonder if we can characterize the class of functions which has a decomposition needed in Algorithms 1 and 2. Furthermore, if a function has this a decomposition, how can we obtain such a decomposition easily. It is worth pointing out that for the case of Algorithm 2, the convex function φ is known in many cases. For instance, in the case of constrained optimization, the function φ is the indicator of the constraint set; in many statistical estimation problems, φ is a penalty function on the parameters; a well-known example of such penalty function is the ℓ_1 penalty, which is used to obtain sparse solutions. Hence, for all practical purposes, the task of characterizing the function class mentioned in Theorems 1 and 2 reduces to characterizing functions which can be decomposed as a difference of a smooth function(g) and a convex function (h). In the next theorem, we characterize the class of of continuously differentiable functions that can be written as a difference of a smooth function and a convex function.

Theorem 6 *Given any continuously differentiable function $f : \mathbb{R}^d \mapsto \mathbb{R}$, the following two properties are equivalent:*

(a) There exists a M-smooth function g, and a convex continuously differentiable function h such that:

$$f(x) = g(x) - h(x) \quad \textit{for all } x \in \mathbb{R}^d.$$

(b) The gradient of the function f satisfies the following inequality:

$$\big\langle \nabla f(x) - \nabla f(y), x - y \big\rangle \leq M\|x - y\|^2 \quad \textit{for all } x, y \in \mathbb{R}^d.$$

Proof We establish the equivalence by proving the circle of implications $(a) \implies (b) \implies (a)$.

Implication (a) $\implies$ (b): For any M-smooth function g, we have the following:

$$\begin{aligned}\big\langle \nabla g(x) - g(y), x - y \big\rangle &\leq \|\nabla g(x) - g(y)\|_2 \times \|x - y\|_2 \\ &\overset{(i)}{\leq} M\|x - y\|_2^2, \quad \text{for all } x, y \in \mathbb{R}^d,\end{aligned} \tag{76}$$

where step (i) follows since the gradient ∇g is M Lipschitz. Next note that the gradient of a differentiable convex function is a monotone operator, and we have that for all $x, y \in \mathbb{R}^d$:

$$\big\langle \nabla h(x) - \nabla h(y), x - y \big\rangle \geq 0. \tag{77}$$

Subtracting equation (77) from equation (76), we obtain the desired upper bound in part (b).

Implication (b) $\implies$ (a): We prove this implication by finding a M–smooth function g and a convex differentiable function h such that $f = g - h$. To this end, we fix any $x_0 \in \mathbb{R}^d$ and consider the following two functions:

$$g(x) := f(x_0) + \big\langle \nabla f(x_0), x - x_0 \big\rangle + \frac{M}{2}\|x - x_0\|_2^2 \tag{78a}$$

$$h(x) := g(x) - f(x). \tag{78b}$$

The function g in definition (78a) is M-smooth by inspection. Since both the functions f and g are continuously differentiable, the function h is continuously differentiable by construction. In order to complete the proof, it suffices to show that the function h is convex. To this end, the first order Taylor series expansion of the function h yields

$$\begin{aligned} h(x) &= h(y) + \big\langle \nabla h(y + t(x-y)), x - y \big\rangle \quad \text{for some } t \ \in [0,1] \\ &= h(y) + \big\langle \nabla h(y), x-y \big\rangle + \big\langle \nabla h(y+t(x-y)) - \nabla h(y), x-y \big\rangle. \end{aligned} \tag{79}$$

Expanding the term $\big\langle \nabla h(y+t(x-y)) - \nabla h(y), x-y \big\rangle$ above yields,

$$\begin{aligned} \big\langle \nabla h(y+t(x-y)) - \nabla h(y), x-y \big\rangle &\overset{(i)}{=} M\|x-y\|_2^2 - \frac{\big\langle \nabla f(y+t(x-y)) - \nabla f(y), t(x-y) \big\rangle}{t} \\ &\overset{(ii)}{\geq} M\|x-y\|_2^2 - Mt\|x-y\|_2^2 \\ &\overset{(iii)}{\geq} 0. \end{aligned}$$

Here step (i) follows by substituting the expression of the function h; step (ii) follows from the gradient inequality of part (b), and step (iii) follows from the inequality $0 \leq t \leq 1$. Since the vectors $x, y \in \mathbb{R}^d$ were arbitrary, the inequality $\big\langle \nabla h(y+t(x-y)) - \nabla h(y), x-y \big\rangle \geq 0$ combined with equation (79) proves the convexity of the function h, thereby proving the claimed result in part (a).

■

Comments: It would be interesting to characterize the class of DC-based functions mentioned in problem (2) when the convex function h is non-differentiable. Indeed, we obtain a larger and more interesting non-differentiable class of functions. It would interesting to see whether Theorem 6 can be suitably generalized in this setting.

References

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

Amir A Ahmadi and Pablo A Parrilo. A complete characterization of the gap between convexity and sos-convexity. *SIAM Journal on Optimization*, 23(2):811–833, 2013.

Amir A Ahmadi, Alex Olshevsky, Pablo A Parrilo, and John N Tsitsiklis. Np-hardness of deciding convexity of quartic polynomials and related problems. *Mathematical Programming*, 137(1-2):453–476, 2013.

Nguyen Thai An and Nguyen Mau Nam. Convergence analysis of a proximal point algorithm for minimizing differences of functions. *Optimization*, 66(1):129–147, 2017.

Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka- Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

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.

Jérôme Bolte, Shoham Sabach, and Marc Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(1-2): 459–494, 2014.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

Yair Carmon, John C Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points i. *arXiv preprint arXiv:1710.11606*, 2017.

Coralia Cartis, Nicholas IM Gould, and Ph L Toint. On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems. *Siam journal on optimization*, 20(6):2833–2852, 2010.

Michel Coste. An introduction to semialgebraic geometry. *RAAG network school*, 145:30, 2002.

Ady Ecker and Allan D Jepson. Polynomial shape from shading. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 145–152. IEEE, 2010.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

Jianqing Fan and Runz Li. Variable selection via non-concave penalized likelihood and its oracle properties. *Jour. Amer. Stat. Ass.*, 96(456):1348–1360, December 2001.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

Jun-Ya Gotoh, Akiko Takeda, and Katsuya Tono. DC formulations and algorithms for sparse optimization problems. *Mathematical Programming*, pages 1–36, 2017.

Philip Hartman. On functions representable as a difference of convex functions. *Pacific Journal of Mathematics*, 9(3):707–713, 1959.

Mingyi Hong, Zhi-Quan Luo, and Meisam Razaviyayn. Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems. *SIAM Journal on Optimization*, 26(1):337–364, 2016.

Reiner Horst, Panos M. Pardalos, and Nguyen Van Thoai. *Introduction to Global Optimization*, volume 9 of *Nonconvex Optimization and its Applications*. Elsevier, 2000.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

Krzysztof Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l'institut Fourier*, volume 48, pages 769–784. Chartres: L'Institut, 1950-, 1998.

Simon Lacoste-Julien. Convergence rate of Frank-Wolfe for non-convex objectives. *arXiv preprint arXiv:1607.00345*, 2016.

Gert R Lanckriet and Bharath K Sriperumbudur. On the convergence of the concave-convex procedure. In *Advances in neural information processing systems*, pages 1759–1767, 2009.

Jason D Lee, Max Simchowitz, Michael I Jordan, and Benjamin Recht. Gradient descent only converges to minimizers. In *Conference on Learning Theory*, pages 1246–1257, 2016.

Guoyin Li and Ting Kei Pong. Calculus of the exponent of Kurdya- Łojasiewicz inequality and its applications to linear convergence of first-order methods. *arXiv preprint arXiv:1602.02915*, 2016.

Yuanzhi Li and Yang Yuan. Convergence analysis of two-layer neural networks with relu activation. In *Advances in Neural Information Processing Systems*, pages 597–607, 2017.

Thomas Lipp and Stephen Boyd. Variations and extension of the convex–concave procedure. *Optimization and Engineering*, 17(2):263–287, 2016.

Po-Ling Loh and Martin J Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. In *Advances in Neural Information Processing Systems*, pages 476–484, 2013.

Yurii Nesterov and Boris T Polyak. Cubic regularization of Newton method and its global performance. *Mathematical Programming*, 108(1):177–205, 2006.

Alexander M Ostrowski. *Solution of equations and systems of equations: Pure and applied mathematics: A Series of monographs and textbooks*, volume 9. Elsevier, 2016.

Ioannis Panageas and Georgios Piliouras. Gradient descent only converges to minimizers: Non-isolated critical points and invariant regions. *arXiv preprint arXiv:1605.00405*, 2016.

Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

T Pham Dinh, HV Ngai, and HA Le Thi. Convergence analysis of the DC algorithm for DC programming with subanalytic data. *preprint*, 2013.

R Tyrrell Rockafellar and Roger J-B Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

Hoang Tuy. DC optimization: theory, methods and algorithms. In *Handbook of global optimization*, pages 149–216. Springer, 1995.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press, Cambridge, UK, 2019.

Shenlong Wang, Alex Schwing, and Raquel Urtasun. Efficient inference of continuous Markov random fields with polynomial potentials. In *Advances in neural information processing systems*, pages 936–944, 2014.

Bo Wen, Xiaojun Chen, and Ting Kei Pong. A proximal difference-of-convex algorithm with extrapolation. *Computational Optimization and Applications*, 69(2):297–324, 2018.

Yangyang Xu and Wotao Yin. A globally convergent algorithm for nonconvex optimization based on block coordinate update. *Journal of Scientific Computing*, 72(2):700–734, 2017.

Bowei Yan, Mingzhang Yin, and Purnamrita Sarkar. Convergence of gradient EM on multi-component mixture of Gaussians. In *Advances in Neural Information Processing Systems*, pages 6959–6969, 2017.

Alan L Yuille and Anand Rangarajan. The concave-convex procedure. *Neural computation*, 15(4):915–936, 2003.

Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999.