A. Further simulation study

In this section, we apply our method to the problem of shape from shading (SFS) reconstruction.

A.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 & 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 2.

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

$$I_{ij} = \frac{\langle L,\, N_{ij} \rangle}{\|N_{ij}\|_2}. \tag{21}$$

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

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

Squaring both sides of equation (21) and substituting the expression for surface normal N_{ij} yields the polynomial equation

$$\big(p_{ij}^2 + q_{ij}^2 + 1\big) I_{ij} - (\ell_1 p_{ij} + \ell_2 q_{ij} + \ell_3)^2 = 0, \tag{22}$$

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 $\big\{z_{ij} \mid (i, j) \in [r] \times [c]\big\}$:

$$\min_{z \in \mathbb{R}^{r \times c}} \quad \underbrace{\Big\{ \sum_{i=1}^{r} \sum_{j=1}^{c} \big((1 + p_{ij}^2 + q_{ij}^2) I_{ij}^2 - (\ell_1 p_{ij} + \ell_2 q_{ij} + \ell_3)^2 \big)^2 \Big\}}_{P(z)}. \tag{23}$$

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 (23) using the backtracking gradient descent algorithm 3. The reconstructed surfaces for *Vase* and *Mozart* are provided in figure 2. We ran 500 iterations of Algorithm 3 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 3 for Problem (23) is parallelizable; hence, the runtime can be much lower than our runtime with a parallel implementation. For sake of completeness we mention the standard backtracking algorithm in Algorithm 3.

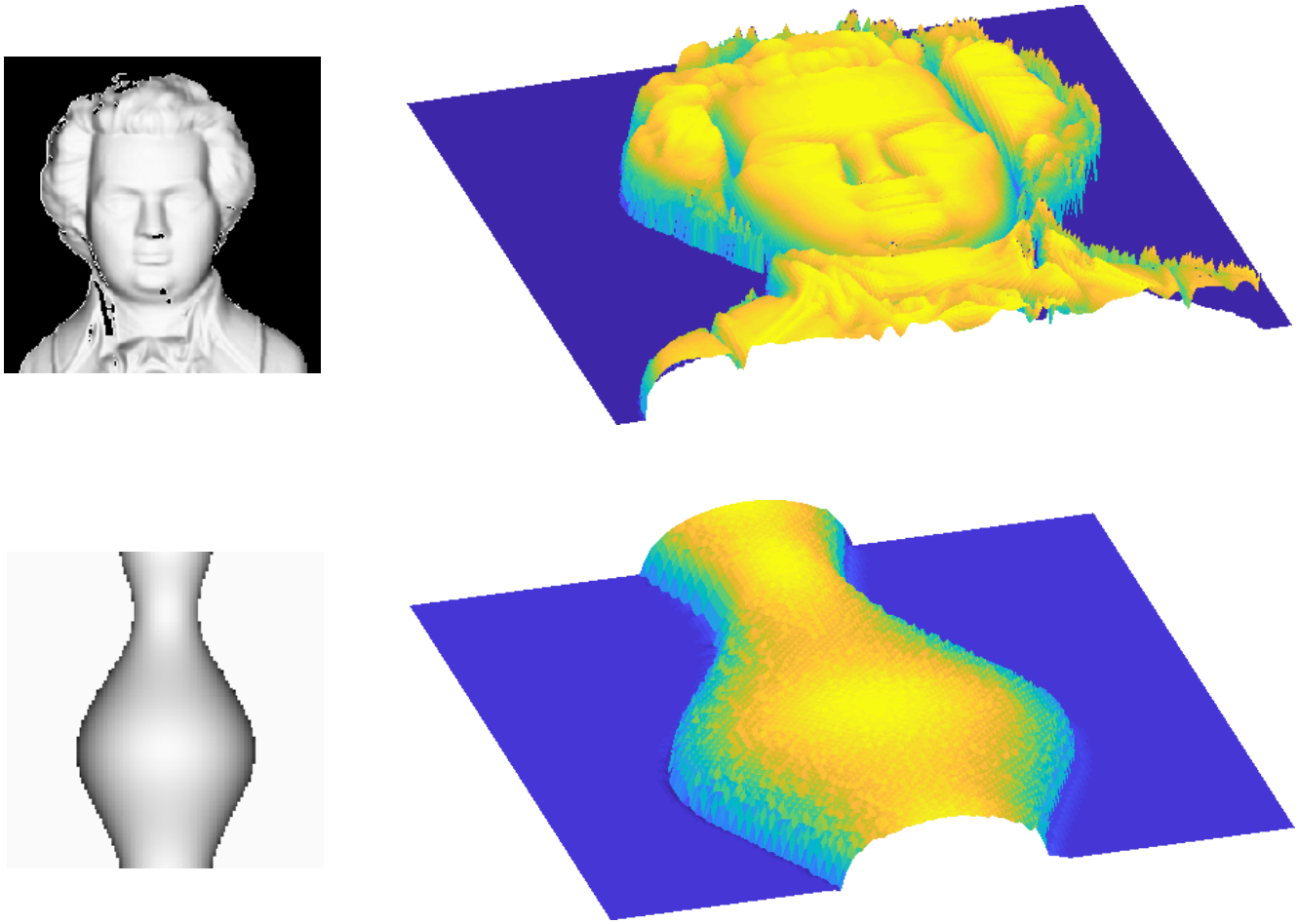

Figure 2. Figure shows 3D shape reconstruction of *Mozart* (first row) and *Vase* (second row) from corresponding 2D images. The gray-scale images in the left column are the 2D input images; the two colored images in the right column are the reconstructed 3D shapes. The 3D shapes are constructed by solving the problem (23) using Algorithm 3.

B. Technical background

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

B.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 $x \in \mathrm{dom}(f)$, the Fréchet subgradient of the function f at point x is defined as

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

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

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

Algorithm 3 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:

$$f\big(x^k - t^k \nabla f(x^k)\big) \leq f(x^k) - \frac{t^k}{2}\|\nabla f(x^k)\|^2. \tag{24}$$

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

Properties: The following properties of Fréchet and limiting sub-differential are provided in Chapter 8 of the book Rockafeller 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 $\mathrm{graph}(\partial_L f) := \big\{(x, u) \in \mathbb{R}^d \times \mathbb{R} \mid u \in \partial_L f(x)\big\}$.

.

B.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 3) or $f = g + \varphi - h$ (Theorems 2 and 4). The following properties are direct consequences of properties of the limiting subdifferential mentioned in Appendix B.1.

- Suppose 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 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 utilizing a previous work by Bolte et al. (2007). In order to facilitate further discussion, we introduce few notations used in the paper (Bolte et al., 2007). 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. 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 3 or Theorem 4, 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

$$\|\nabla f(x)\|_2 = \|\partial_L(-f)(x)\|_2 \overset{(i)}{\geq} m_{(-f)}(x), \tag{25}$$

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

$$\|\nabla f(x)\|_2 = \|\partial_L f(x)\|_2 \overset{(i)}{\geq} m_f(x), \tag{26}$$

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

B.3. Instances of sub-analytic functions

In Appendix B.2, we proved that continuous sub-analytic functions satisfy Assumption KL, and in those cases,—by Theorems 3 and 4—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 $\mathrm{graph}(f) := \big\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : f(x) = y\big\}$, 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 (Facchinei & 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) Pointwise 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*.

C. 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, Corollary 1 and Proposition 1.

C.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

$$x^k \in \operatorname{int}(\mathcal{C}) \quad \text{and} \quad f(x^{k+1}) \leq f(x^k) - \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2 \quad \text{for all } k = 0, 1, 2, \ldots. \tag{27}$$

See Appendix C.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 $\{f(x^k)\}_{k\geq 0}$ is convergent. Since $f^* := \min_{x\in\mathcal{C}} f(x)$ is finite by assumption, and $x^k \in \operatorname{int}(\mathcal{C})$ for all $k \geq 0$ by Lemma 4, the sequence $\{f(x^k)\}_{k\geq 0}$ is bounded below. For any non-stationary x^k, inequality (27) 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 $\{x^k\}_{k\geq 0}$ must be stationary. Consider a subsequence $\{x^{k_j}\}_{j\geq 0}$ of $\{x^k\}_{k\geq 0}$ such that $x^{k_j} \to \bar{x}$, and let $\{u^{k_j}\}_{j\geq 0}$ be the associated sequence of subgradients. It suffices to exhibit a sub-gradient $\bar{u} \in \partial h(\bar{x})$ such that $\nabla g(\bar{x}) - \bar{u} = 0$.

Since the sequence $\{x^{k_j}\}_{j\geq 0}$ 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 B.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

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

where step (i) follows from equation (27). Rearranging yields the claimed bound (3) on the averaged squared gradient.

C.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

$$g(x) \leq g(x^k) + \langle \nabla g(x^k),\, x - x^k \rangle + \frac{M_g}{2}\|x - x^k\|_2^2 \tag{28a}$$

$$h(x) \geq h(x^k) + \langle u^k,\, x - x^k \rangle. \tag{28b}$$

Combining inequalities (28a) and (28b) yield

$$f(x) = g(x) - h(x) \leq f(x^k) + \langle \nabla g(x^k) - u^k,\, x - x^k \rangle + \frac{M_g}{2}\|x - x^k\|_2^2. \tag{29}$$

Substituting $x = x^{k+1} := x^k - \alpha\big(\nabla g(x^k) - u^k\big)$ in equation (29) and simplifying yields

$$\begin{aligned} f(x^k) - f(x^{k+1}) \geq \big(\frac{1}{\alpha} - \frac{M_g}{2}\big)\|x^{k+1} - x^k\|_2^2 = \alpha\big(1 - \frac{\alpha M_g}{2}\big)\|\nabla g(x^k) - u^k\|_2^2 \\ \overset{(i)}{\geq} \frac{\alpha}{2}\|\nabla f(x^k)\|_2^2, \end{aligned}$$

where inequality (i) follows from the upper bound $\alpha \leq \frac{1}{M_g}$. 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 $f(x^{k+1}) \leq f(x^k) \leq f(x^0)$, which ensures that $x^{k+1} \in \mathcal{L}(f(x^0)) \subset \text{int}(\mathcal{C})$, as claimed.

C.2. Proof of Corollary 1

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

$$\arg\min_{x \in \mathbb{R}^d} \Big\{ \frac{1}{2}\|x - y\|_2^2 - \alpha g(x) + \alpha h(x) \Big\}. \tag{30}$$

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 (30) 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 (30), 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 injectivitivty 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.

C.3. 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

$$q(x, x^k) := g(x) - h(x^k) - \langle \nabla h(x^k),\, x - x^k \rangle. \tag{31}$$

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

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\overset{(i)}{\geq} q(x^k, x^k) - q(x^{k+1}, x^k) \\ &\overset{(ii)}{\geq} q(x^k, x^k) - q(y^{k+1}, x^k) \\ &\overset{(iii)}{\geq} \frac{1}{2M_g}\|\nabla f(x^k)\|_2^2. \end{aligned} \tag{32}$$

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 (32) 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 (32) and re-arranging yields inequality (8a). Finally, we turn to the proof of inequality (8b) under the additional strong convexity condition. Under this condition, the map $x \mapsto q(x, x^k)$ in equation (31) is μ-strongly convex, so that

$$f(x^k) - f(x^{k+1}) \geq q(x^k, x^k) - q(x^{k+1}, x^k) \overset{(i)}{\geq} \frac{\mu}{2}\|x^k - x^{k+1}\|_2^2, \tag{33}$$

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 (8b). Finally, let us reiterate that bounds similar to (8b) are known in the literature; see the paper (Lanckriet & Sriperumbudur, 2009) for example. We provide the proof of bound (8b) for completeness.

D. 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

$$x^{k+1} = x^k - \alpha(\nabla g(x^k) + v^{k+1} - u^k), \qquad \text{and} \tag{34a}$$

$$f(x^k) - f(x^{k+1}) \geq \frac{1}{2\alpha} \|x^k - x^{k+1}\|_2^2, \tag{34b}$$

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

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

See Appendix D.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 $k = 0, 1, 2, \ldots$ In that case, inequality (34b) 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 & 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 B.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 B.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 (10a): Next, we establish the claimed bound (10a) on the averaged squared successive difference. Recalling that $f^* := \min_{x \in \mathbb{R}^d} f(x)$ is finite, we have

$$\begin{aligned} f(x^0) - f^* \geq f(x^0) - f(x^{k+1}) &= \sum_{j=0}^{k} f(x^j) - f(x^{j+1}) \\ &\overset{(i)}{\geq} \frac{1}{2\alpha} \sum_{j=0}^{k} \|x^j - x^{j+1}\|_2^2 \\ &= \frac{(k+1)}{2\alpha} \text{Avg}\left(\|x^k - x^{k+1}\|_2^2\right), \end{aligned} \tag{35}$$

where step (i) follows from equation (34b). Rearranging the last inequality yields the claimed bound (10a) on the averaged squared successive difference.

Establishing the bound (10b): In order to establish the bound (10b) on the averaged squared gradient, we start by establishing the following upper bound on the gradient-norm $\|\nabla f(x^{k+1})\|_2$:

$$\|\nabla f(x^{k+1})\|_2 \leq \big(M_g + M_h + \frac{1}{\alpha}\big)\|x^k - x^{k+1}\|_2. \tag{36}$$

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 (36) and (35), we obtain the desired bound (10b).

D.1. Proof of Lemma 5

Here we prove the claims of Lemma 5.

Establishing update equation (34a): Recalling the convex majorant defined in equation (29), we define a convex majorant $q(\cdot, x^k)$ of the function f as follows:

$$q(x, x^k) = g(x^k) - h(x^k) + \langle \nabla g(x^k) - u^k,\, x - x^k \rangle + \frac{1}{2\alpha}\|x - x^k\|_2^2 + \varphi(x), \tag{37}$$

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 $\text{prox}^{\varphi}_{1/\alpha}\big(x^k - \alpha(\nabla g(x^k) - u^k)\big)$, 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:

$$\nabla g(x^k) - u^k + v^{k+1} + \frac{1}{\alpha}\big(x^{k+1} - x^k\big) = 0. \tag{38}$$

Rewriting the above equation yields the update equation (34a).

Establishing the descent step (34b): Note that

$$\begin{aligned} f(x^k) - q(x^{k+1}, x^k) &\overset{(i)}{\geq} g(x^k) - h(x^k) + \varphi(x^{k+1}) + \langle v^{k+1},\, x^k - x^{k+1}\rangle - q(x^{k+1}, x^k) \\ &\overset{(ii)}{\geq} \langle \nabla g(x^k) - u^k + v^{k+1},\, x^k - x^{k+1}\rangle - \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\ &\overset{(iii)}{\geq} \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2. \end{aligned} \tag{39}$$

Here step (i) follows from the convexity of the function φ; step (ii) follows by substituting $q(x^{k+1}, x^k)$ from equation (37). 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 (38). Finally, recall that the function $x \mapsto q(x, x^k)$ is a majorant for the function f, and we deduce that

$$\begin{aligned} f(x^k) - f(x^{k+1}) &\geq f(x^k) - q(x^{k+1}, x^k) \\ &\geq \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2. \end{aligned} \tag{40}$$

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 (34b) 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

$$\liminf_{j\to\infty}\ \varphi(x^{k_j+1}) \geq \varphi(\bar{x}). \tag{41}$$

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

$$\begin{aligned} \limsup_{j\to\infty}\ \varphi(x^{k_j+1}) &\overset{(i)}{\leq} \varphi(\bar{x}) + \limsup_{j\to\infty}\ \Big(\langle \bar{x} - x^{k_j},\, \nabla g(x^{k_j}) - u^{k_j}\rangle + \frac{1}{2\alpha}\|x^{k_j} - \bar{x}\|_2^2\Big) \\ &\overset{(ii)}{=} \varphi(\bar{x}). \end{aligned} \tag{42}$$

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 (41) and (42) we obtain the claimed result.

Boundedness of the sequence $\big\{\nabla g(x^{k_j}) - u^{k_j}\big\}_{j\geq 0}$: In order to prove the boundedness of the sequence $\big\{\nabla g(x^{k_j}) - u^{k_j}\big\}_{j\geq 0}$, it suffices to show that the gradient sequence $\big\{\nabla g(x^{k_j})\big\}_{j\geq 0}$ and the sub-gradient sequence $\big\{u^{k_j}\big\}_{j\geq 0}$ 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 $\big\{\nabla g(x^{k_j})\big\}_{j\geq 0}$ 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 & Wets, 2009), we conclude that the subgradient sequence $\big\{u^{k_j}\big\}_{j\geq 0}$ bounded.

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 3 and 4. We begin by stating an auxiliary lemma that underlies the proofs of Theorems 3 and 4.

Lemma 6. Under assumptions of either Theorem 3 or Theorem 4, 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 3

Now we prove Theorem 3 using Lemma 6.

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

$$\begin{aligned}
\big(f(x^k) - \bar{f}\big)^{1-\gamma\theta} - \big(f(x^{k+1}) - \bar{f}\big)^{1-\gamma\theta} &\geq \big(1 - \gamma\theta\big)\big(f(x^k) - \bar{f}\big)^{-\gamma\theta}\big(f(x^k) - f(x^{k+1})\big) \\
&\overset{(i)}{\geq} \big(1 - \gamma\theta\big)\big(|f(x^k) - \bar{f}|\big)^{-\gamma\theta} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\
&\overset{(ii)}{\geq} \frac{(1 - \gamma\theta)}{C\|\nabla f(x^k)\|_2^\gamma} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\
&\overset{(iii)}{=} \frac{(1 - \gamma\theta)}{2C\alpha^{1-\gamma}}\|x^k - x^{k+1}\|_2^{2-\gamma}. \qquad (43)
\end{aligned}$$

Here inequality (i) follows from the descent property in equation (27) and from the fact that $f(x^k) \downarrow \bar{f}$. Inequality (ii) follows from Lemma 6, 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 (43) 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 $\big(\|\nabla f(x^k)\|_2\big)$: Rewriting equation (43), we have the following:

$$\begin{aligned}
C_\gamma &:= \sum_{\ell=0}^{k_1} \frac{(1 - \gamma\theta)}{2C\alpha^{1-\gamma}}\|x^\ell - x^{\ell+1}\|_2^{2-\gamma} + \big(f\big(x^{k_1}\big) - \bar{f}\big)^{(1-\gamma\theta)} \\
&\overset{(i)}{\geq} \sum_{\ell=0}^{k-1} \frac{(1 - \gamma\theta)}{2C\alpha^{1-\gamma}}\|x^\ell - x^{\ell+1}\|_2^{2-\gamma} \\
&= \frac{k(1 - \gamma\theta)}{2C\alpha^{1-\gamma}}\text{Avg}\left(\|x^k - x^{k+1}\|_2^{2-\gamma}\right), \qquad (44)
\end{aligned}$$

where step (i) above follows from equation (43), and 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 (44), 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 $\left(\|\nabla f(x^k)\|_2\right)$: Since we proved that the sequence $\left\{x^k\right\}_{k\geq 0}$ is convergent to the point $\bar{x}$, we have that the parameter θ in Lemma 6 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 (44) yields,

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

where $c_2 = \frac{1}{\alpha}\left(\frac{2CC_\gamma\alpha^{1-\gamma\theta}}{1-\gamma\theta}\right)^r$ with $\gamma = \frac{2r-1}{r}$, and GAvg $\left(\|x^k - x^{k+1}\|_2^{2-\gamma}\right) := \prod_{\ell=0}^{k-1}\left(\|x^\ell - x^{\ell+1}\|_2\right)^{\frac{1}{k}}$, the geometric average of the sequence $\left\{\|x^\ell - x^{\ell+1}\|_2\right\}_{l=0}^{k-1}$. Here step (i) above follows from arithmetic-geometric mean (AM/GM) inequality; step (ii) follows from the bound in equation (44) and from the fact that $\gamma = \frac{2r-1}{r}$. Finally, note that the last equality holds trivially for iteration $k \leq k_1$ with the given choice of constant c_2.

E.2. Proof of Theorem 4

The proof of Theorem 4 builds on the techniques used in the proof of Theorem 3 but requires additional technical care due to the presence of possibly non-continuous function φ.

Convergence of the sequence $\left\{x^k\right\}_{k\geq 0}$: The proof of Theorem 4 has two steps. First, we prove a descent condition similar to equation (43). We then leverage this descent condition and weighted AM-GM inequality to obtain the desired result.

Step 1: Following the proof of Theorem 3, we prove the convergence of the sequence $\left\{x^k\right\}_{k\geq 0}$ by showing that the sequence $\left\{x^k\right\}_{k\geq 0}$ has finite length property. First, note that for scalars $0 \leq \theta < 1$ and $0 < \gamma < \frac{1}{\theta}$, the function $t \mapsto t^{1-\gamma\theta}$ is concave. Consequently, for iteration $k \geq k_1$, from Lemma 6 we have

$$\begin{aligned}
\left(f(x^k) - \bar{f}\right)^{1-\gamma\theta} - \left(f(x^{k+1}) - \bar{f}\right)^{1-\gamma\theta} &\geq \left(1-\gamma\theta\right)\left(f(x^k) - \bar{f}\right)^{-\gamma\theta}\left(f(x^k) - f(x^{k+1})\right) \\
&\overset{(i)}{\geq} \left(1-\gamma\theta\right)\left(|f(x^k) - \bar{f}|\right)^{-\gamma\theta} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2 \\
&\overset{(ii)}{\geq} \frac{(1-\gamma\theta)}{C\|\nabla f(x^k)\|_2^\gamma} \times \frac{1}{2\alpha}\|x^k - x^{k+1}\|_2^2. \qquad (45)
\end{aligned}$$

Here step (i) follows from the descent property in equation (40) and from the fact that $f(x^k) \downarrow \bar{f}$; step (ii) follows from Lemma 6. 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 $\left\{x^k\right\}_{k\geq 0}$ 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:

$$\|\nabla g(x^k) - \nabla h(x^k) + v^k\|_2 \leq \left(M_{g-h} + \frac{1}{\alpha}\right)\|x^k - x^{k-1}\|_2. \qquad (46)$$

Combining the last inequality with inequality (45) yields the following descent property

$$\left(f(x^k) - \bar{f}\right)^{1-\gamma\theta} - \left(f(x^{k+1}) - \bar{f}\right)^{1-\gamma\theta} \geq \frac{(1-\gamma\theta)}{2\alpha C\left(M_{g-h} + \frac{1}{\alpha}\right)^\gamma} \times \frac{\|x^k - x^{k+1}\|_2^2}{\|x^k - x^{k-1}\|_2^\gamma}. \qquad (47)$$

Step 2: We now leverage the descent condition obtained from step 1 to prove finite length property of the sequence $\{x^k\}_{k\geq 0}$. 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 (47) as

$$\Delta_\gamma^k \|x^{k-1}-x^k\|_2^\gamma \geq \|x^k - x^{k+1}\|_2^2. \tag{48}$$

Combining equation (48) with the weighted AM-GM inequality, we obtain

$$\begin{aligned}\big(1+\frac{\gamma}{2-\gamma}\big)\times\sum_{j=k_1+1}^{k}\|x^j-x^{j+1}\|_2^{2-\gamma} &\overset{(i)}{\leq} \big(1+\frac{\gamma}{2-\gamma}\big)\times\sum_{k=k_1+1}^{k}\Big(\sqrt{\Delta_\gamma^j\,\|x^{j-1}-x^j\|_2^\gamma}\Big)^{\frac{2-\gamma}{2}}\\ &\overset{(ii)}{\leq}\sum_{j=k_1+1}^{k}\big(\Delta_\gamma^j+\frac{\gamma}{2-\gamma}\|x^{j-1}-x^j\|_2^{2-\gamma}\big)\\ &\overset{(iii)}{\leq} C_3\big(f(x^{k_1})-\bar{f}\big)^{1-\gamma\theta}+\sum_{j=k_1+1}^{k}\frac{\gamma}{2-\gamma}\|x^{j-1}-x^j\|_2^{2-\gamma}.\end{aligned} \tag{49}$$

Here step (i) follows from equation (48), 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 (49) 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 (49), we have for all $k\geq k_1+2$

$$\begin{aligned}\sum_{j=k_1+1}^{k-1}\|x^j-x^{j+1}\|_2^{2-\gamma} &\leq C_3\big(f(x^{k_1})-\bar{f}\big)^{1-\gamma\theta}+\frac{\gamma}{2-\gamma}\|x^{k_1}-x^{k_1+1}\|_2^{2-\gamma}-\big(1+\frac{\gamma}{2-\gamma}\big)\|x^k-x^{k+1}\|_2^{2-\gamma}\\ &\leq C_3\big(f(x^{k_1})-\bar{f}\big)^{1-\gamma\theta}+\frac{\gamma}{2-\gamma}\|x^{k_1}-x^{k_1+1}\|_2^{2-\gamma}<\infty.\end{aligned} \tag{50}$$

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 3 and using the inequality (50) and upper bound (46).

E.3. Proof of Lemma 6

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 $\left\{x^k\right\}_{k \geq 0}$ 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 $\bar{\mathcal{X}} = \{\bar{x}\}$.

F. Proofs of Corollaries:

In this appendix, we collect the proofs of various Corollaries that appear in Section 5, which include Corollaries 2 and 3.

F.1. Proof of Corollary 2

First, note that in order to apply Theorem 1 and Theorem 3 to Corollary 2, 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 (14) is continuous sub-analytic (see Appendix B.2). For proving sub-analyticity, we heavily use the properties mentioned in Appendix B.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 $\mu \mapsto \Psi\big(y_i - \langle z_i, \mu \rangle\big)$ 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 $\left\{(z_i, y_i)\right\}_{i=1}^{n}$ 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 3 are applicable for problem (14). Convergence of the sequence $\left\{\mu^k\right\}_{k \geq 0}$ to a point $\bar{\mu}$ and the convergence rate of gradient norms follows from Theorem 3, 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 1, it follows that with random initializations, Algorithm 1 avoids strict saddle points almost surely.

F.2. Proof of Corollary 3

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 $\left\{x^k\right\}_{k \geq 0}$—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 (18), 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 4 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 $\mathrm{sign}(x_{(i)}) = \mathrm{sign}(\bar{x}_{(i)})$ 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 3

$$\mathcal{P}(\bar{x}) := \min_{x\in\mathbb{R}^d} g(x) + \lambda\varphi(x) - \lambda\langle\nabla h(\bar{x}),\, x\rangle. \tag{51}$$

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

$$\nabla g(\bar{x}) + \lambda\bar{u} - \lambda\nabla h(\bar{x}) = 0 \quad \text{and} \quad \nabla g(x^*) + \lambda u^* - \lambda\nabla h(x^*) = 0, \tag{52}$$

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 (52) 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 (51); this contradicts the assumption that problem (51) 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 & 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 minimum of finitely many functions as follows:

$$f(x) = \min_{a\in\mathcal{A}}\{\|y - Bx\|_2^2 + \lambda\|x\|_1 - \lambda a^\top x\}, \tag{53}$$

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 & 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 4 part(b), we conclude that:

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

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 (50) which yields

$$\sum_{\ell=k_1+1}^{\infty}\|x^\ell - x^{\ell+1}\|_2 \leq \|x^{k_1} - x^{k_1+1}\|_2 + C_3\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}}, \tag{54}$$

for some constant C_3. From Lemma 6 and equation (36), we have

$$\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \leq C\|\nabla f(x^{k_1})\|_2 \leq C(M + M_h + \frac{1}{\alpha})\|x^{k_1} - x^{k_1-1}\|_2. \tag{55}$$

Combining equations (54) and (55) we have

$$\begin{aligned}\sum_{\ell=k_1}^{\infty} \|x^{\ell} - x^{\ell+1}\|_2 &\leq 2\|x^{k_1} - x^{k_1+1}\|_2 + C_3\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \\ &\overset{(i)}{\leq} 2\|x^{k_1} - x^{k_1+1}\|_2 + CC_3(M + M_h + \frac{1}{\alpha})\|x^{k_1} - x^{k_1-1}\|_2 \\ &\overset{(ii)}{\leq} \bar{C}\|x^{k_1} - x^{k_1-1}\|_2, \end{aligned} \tag{56}$$

where $\bar{C}$ is a constant depending on M, M_h, α, C_3 and C, and step (i) above follows from equation (55). 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 (56). Denote $e_k = \sum_{\ell=k}^{\infty} \|x^{\ell} - x^{\ell+1}\|_2$. Then equation (56) provides the following recursion

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

Simple inspection of proof of Theorem 4 and derivations so far ensure that we can derive the equations (54) and (55) 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 (56): 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 (40) we have:

$$\big(f(x^{k_1}) - \bar{f}\big)^{\frac{1}{2}} \geq \big(f(x^{k_1}) - f(x^{k_1+1})\big)^{\frac{1}{2}} \geq \frac{1}{\sqrt{2\alpha}}\|x^{k_1} - x^{k_1+1}\|_2. \tag{57}$$

Combining equations (57) and (55) 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 $k \geq k_1$.