Submitted to the Annals of Statistics

ANALYSIS OF GENERALIZED BREGMAN SURROGATE ALGORITHMS FOR NONSMOOTH NONCONVEX STATISTICAL LEARNING

BY YIYUAN SHE, ZHIFENG WANG AND JIUWU JIN

Department of Statistics, Florida State University

Modern statistical applications often involve minimizing an objective function that may be nonsmooth and/or nonconvex. This paper focuses on a broad Bregman-surrogate algorithm framework including the local linear approximation, mirror descent, iterative thresholding, DC programming and many others as particular instances. The recharacterization via generalized Bregman functions enables us to construct suitable error measures and establish global convergence rates for nonconvex and nonsmooth objectives in possibly high dimensions. For sparse learning problems with a composite objective, under some regularity conditions, the obtained estimators as the surrogate's fixed points, though not necessarily local minimizers, enjoy provable statistical guarantees, and the sequence of iterates can be shown to approach the statistical truth within the desired accuracy geometrically fast. The paper also studies how to design adaptive momentum based accelerations without assuming convexity or smoothness by carefully controlling stepsize and relaxation parameters.

1. Introduction. Many statistical learning problems can be formulated as minimizing a certain objective function. In shrinkage estimation, the objective can often be represented as the sum of a loss function and a penalty function, neither of which is necessarily smooth or convex. For example, when the number of variables is much larger than the number of observations ($p \gg n$), sparsity-inducing penalties come into play and result in nondifferentiability. Furthermore, many popular penalties are nonconvex [22, 19, 65], making the computation and analysis more challenging. Although in low dimensions there are ways to tackle nonsmooth nonconvex optimization, statisticians often prefer easy-to-implement algorithms that scale well in big data applications. Therefore, first-order methods, gradient-descent type algorithms in particular, have recently attracted a great deal of attention due to their lower complexity per iteration and better numerical stability than Newton-type algorithms.

In this work, we study a class of algorithms in a *Bregman surrogate* framework. The idea is that instead of solving the original problem $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$, one constructs a surrogate function

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^-), \tag{1}$$

and generates a sequence of iterates according to

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}). \tag{2}$$

The generalized Bregman function $\boldsymbol{\Delta}_\psi$ will be rigourously defined in Section 2.1, and we will call g a (generalized) Bregman surrogate. Note that $\boldsymbol{\Delta}_\psi$ is not necessarily the standard Bregman divergence [9] because we do not restrict ψ to be smooth or strictly convex or even convex. Bregman divergence does not seem to have been widely used in the statistics community, but see [64]. The generalized Bregman surrogate framework has a close connection

MSC2020 subject classifications: Primary 90C26, 49J52, 68Q25.

Keywords and phrases: nonconvex optimization, nonsmooth optimization, MM algorithms, Bregman divergence, statistical algorithmic analysis, momentum-based acceleration.

to the majorization-minimization (MM) principle [28, 29]. But the surrogate here as a function of $\boldsymbol{\beta}$ matches $f(\boldsymbol{\beta})$ to a higher order when $\boldsymbol{\beta}^-$ is set to $\boldsymbol{\beta}$ (cf. Lemma 4) and we do not always invoke the majorization condition $g(\boldsymbol{\beta};\boldsymbol{\beta}^-) \geq f(\boldsymbol{\beta})$; the benefits will be seen in step size control and acceleration.

A variety of algorithms can be recharacterized by Bregman surrogates, including DC programming [55], local linear approximation (LLA) [67] and iterative thresholding [8, 47]. In contrast to the large body of literature in convex optimization, little research has been done on the rate of convergence of nonconvex optimization algorithms when $p > n$, and there is a lack of universal methodologies. Instead of proving local convergence results for some carefully chosen initial points, this work aims to establish *global* convergence rates regardless of the specific choice of the starting point, where a crucial element is the error measure. We will see that the most natural measures are unsurprisingly problem-dependent, but can be conveniently constructed via generalized Bregman functions.

Another perhaps more intriguing question to statisticians is how the statistical accuracy improves or deteriorates as the cycles progress, and whether the finally obtained estimators can enjoy provable guarantees in a statistical sense. See, for example, [1, 20, 63]; in particular, [36], one of the main motivations of our work, showed that for a composite objective composed of a loss and a regularizer that enforces sparsity, the sequence of iterates $\boldsymbol{\beta}^{(t)}$ generated by gradient-descent type algorithms can approach a minimizer $\boldsymbol{\beta}^o$ at a linear rate even when $p > n$, if the problem under consideration satisfies some regularity conditions. This article reveals broader conclusions when using generalized Bregman surrogate algorithms in the composite setting: the more straightforward *statistical error* between the t-th iterate $\boldsymbol{\beta}^{(t)}$ and the statistical truth $\boldsymbol{\beta}^*$ enjoys fast convergence, and the convergent fixed points, though not necessarily local minimizers, let alone global minimizers, possess the desired statistical accuracy in a minimax sense. The studies support the practice of avoiding unnecessary overoptimization in high-dimensional sparse learning tasks. Our theory will make heavy use of the calculus of generalized Bregman functions—in fact, the proofs become readily on hand with some nice properties of $\boldsymbol{\Delta}$ established. Again, a wise choice of the discrepancy measure can facilitate theoretical analysis and lead to less restrictive regularity conditions.

Finally, we would like to study and extend Nesterov's first and second accelerations [39, 40]. Accelerated gradient algorithms [4, 57, 32] have lately gained popularity in high-dimensional convex programming because they can attain the optimal rates of convergence among first-order methods. However, since convexity is indispensable to these theories, how to adapt the momentum techniques to nonsmooth nonconvex programming is largely unknown. Ghadimi and Lan [24] studied how to accelerate gradient descent type algorithms when the objective function is nonconvex but strongly smooth; the obtained convergence rate is of the same order as gradient descent for nonconvex problems. We are interested in more general Bregman surrogates with a possible lack of smoothness and convexity, most notably in high-dimensional nonconvex sparse learning. This work will come up with two momentum-based schemes to accelerate Bregman-surrogate algorithms by carefully controlling the sequences of relaxation parameters and step sizes.

Overall, this paper aims to provide a universal tool of generalized Bregman functions in the interplay between optimization and statistics, and to demonstrate its active roles in constructing error measures, formulating less restrictive regularity conditions, characterizing strong convexity, deriving the so-called basic inequalities in nonasymptotic statistical analysis, devising line search and momentum-based updates, and so on. The rest of this paper is organized as follows. In Section 2, we introduce the generalized Bregman surrogate framework and present some examples. Section 3 gives the main theoretical results on computational accuracy and statistical accuracy. Section 4 proposes and analyzes two acceleration schemes. We conclude in Section 5. Simulation studies and all technical details are provided in the Appendices.

Notation. Throughout the paper, we use C, c to denote positive constants. They are not necessarily the same at each occurrence. The class of continuously differentiable functions is denoted by $\mathcal{C}^1$. Given any matrix $\boldsymbol{A}$, we denote its (i,j)-th element by A_{ij}. The spectral norm and the Frobenius norm of $\boldsymbol{A}$ are denoted by $\|\boldsymbol{A}\|_2$ and $\|\boldsymbol{A}\|_F$, respectively. The Hadamard product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ of the same dimension is denoted by $\boldsymbol{A} \circ \boldsymbol{B}$ and their inner product is $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = tr\{\boldsymbol{A}^\top \boldsymbol{B}\}$. If $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite, we also write $\boldsymbol{A} \succeq \boldsymbol{B}$. Let $[p] := \{1, \cdots, p\}$. Given $\mathcal{J} \subset [p]$, we use $\boldsymbol{A}_{\mathcal{J}}$ to denote the submatrix of $\boldsymbol{A}$ formed by the columns indexed by $\mathcal{J}$. Given a set $A \subset \mathbb{R}^n$, we use A°, $\mathrm{ri}(A)$, $\overline{A}$ to denote its interior, relative interior, and closure, respectively [45]. When f is an extended real-valued function from $D \subset \mathbb{R}^p$ to $\mathbb{R} \cup \{+\infty\}$, its effective domain is defined as $\mathrm{dom}(f) = \{\boldsymbol{\beta} \in \mathbb{R}^p : f(\boldsymbol{\beta}) < +\infty\}$. Let $\mathbb{R}_+ = [0, +\infty)$.

2. Basics of generalized Bregman surrogates.

2.1. *Generalized Bregman functions.* Bregman divergence [9], typically defined for continuously differentiable and strictly convex functions, plays an important role in convex analysis. An extension of it based on "right-hand" Gateaux differentials helps to handle nonsmooth nonconvex optimization problems. We begin with one-sided directional derivative.

DEFINITION 1. *Let $\psi : D \subset \mathbb{R}^p \to \mathbb{R}$ be a function. The one-sided directional derivative of ψ at $\boldsymbol{\beta} \in D$ with increment $\boldsymbol{h}$ is defined as*

$$\delta\psi(\boldsymbol{\beta}; \boldsymbol{h}) = \lim_{\epsilon \to 0+} \frac{\psi(\boldsymbol{\beta} + \epsilon \boldsymbol{h}) - \psi(\boldsymbol{\beta})}{\epsilon}, \tag{3}$$

provided $\boldsymbol{h}$ is admissible in the sense that $\boldsymbol{\beta} + \epsilon \boldsymbol{h} \in D$ for sufficiently small $\epsilon : 0 < \epsilon < \epsilon_0$. When $\psi : D \to \mathbb{R}^n$ is a vector function, $\delta\psi$ is defined componentwise.

In the following, ψ is called (one-sided) directionally differentiable at $\boldsymbol{\beta}$ if $\delta\psi(\boldsymbol{\beta}; \boldsymbol{h})$ as defined in (3) exists *and* is finite for all admissible $\boldsymbol{h}$, and if this holds for all $\boldsymbol{\beta} \in D$, we say that ψ is directionally differentiable.

When $a > 0$, $\delta\psi(\boldsymbol{\beta}; a\boldsymbol{h}) = a\delta\psi(\boldsymbol{\beta}; \boldsymbol{h})$, but $\delta\psi$ is not necessarily a linear operator with respect to $\boldsymbol{h}$. Definition 1 is a relaxed version of the standard Gateaux differential which studies the limit when $\epsilon \to 0$. In high-dimensional sparse problems where nonsmooth regularizers and/or losses are widely used, (3) is more convenient and useful.

DEFINITION 2 (**Generalized Bregman Function (GBF)**). *The generalized Bregman function associated with a function ψ is defined by*

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \delta\psi(\boldsymbol{\gamma}; \boldsymbol{\beta} - \boldsymbol{\gamma}), \tag{4}$$

assuming $\boldsymbol{\beta}, \boldsymbol{\gamma} \in dom(\psi)$ and $\delta\psi(\boldsymbol{\gamma}; \boldsymbol{\beta} - \boldsymbol{\gamma})$ is meaningful and finite. In particular, when ψ is differentiable and strictly convex, the generalized Bregman function $\boldsymbol{\Delta}_\psi$ becomes the standard Bregman divergence:

$$\mathbf{D}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \langle \nabla\psi(\boldsymbol{\gamma}), \boldsymbol{\beta} - \boldsymbol{\gamma} \rangle. \tag{5}$$

When ψ is a vector function, a vector version of $\boldsymbol{\Delta}$ is defined componentwise.

When $\nabla\psi$ exists at $\boldsymbol{\beta}$, $\delta\psi(\boldsymbol{\beta}, \boldsymbol{h})$ reduces to $\langle \nabla\psi(\boldsymbol{\beta}), \boldsymbol{h} \rangle$, which is linear in $\boldsymbol{h}$. So if ψ is the restriction of a function $\varphi \in \mathcal{C}^1$ to a convex set, $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \boldsymbol{\Delta}_\varphi(\boldsymbol{\beta}, \boldsymbol{\gamma})$ for all $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathrm{dom}(\psi)$. For simplicity, all functions in our paper are assumed to be defined on a *whole* vector space ($\mathbb{R}^p$, typically) unless otherwise mentioned, although most results can be formulated in the case of extended real-valued functions under the convexity of their effective domains.

The generalized Bregman $\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\gamma})$ can be seen as the difference between the function ψ and its radial approximations made at $\boldsymbol{\gamma}$. A simple but important example is $\mathbf{D}_2(\boldsymbol{\beta},\boldsymbol{\gamma}) := \mathbf{D}_{\|\cdot\|_2^2/2}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \|\boldsymbol{\beta}-\boldsymbol{\gamma}\|_2^2/2$. In general, $\boldsymbol{\Delta}_\psi$ or $\mathbf{D}_\psi$ may not be symmetric. The following symmetrized version turns out to be useful:

$$\bar{\boldsymbol{\Delta}}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) := \frac{1}{2}(\boldsymbol{\Delta}_\psi + \grave{\boldsymbol{\Delta}}_\psi)(\boldsymbol{\beta},\boldsymbol{\gamma}) = \frac{1}{2}\{\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\gamma},\boldsymbol{\beta})\}, \tag{6}$$

where $\grave{\boldsymbol{\Delta}}(\boldsymbol{\beta},\boldsymbol{\gamma})$ denotes $\boldsymbol{\Delta}(\boldsymbol{\gamma},\boldsymbol{\beta})$. If ψ is smooth, $\bar{\boldsymbol{\Delta}}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \langle \nabla\psi(\boldsymbol{\beta}) - \nabla\psi(\boldsymbol{\gamma}), \boldsymbol{\beta}-\boldsymbol{\gamma}\rangle$.

To simplify the notation, we use $\boldsymbol{\Delta}_\psi \geq \boldsymbol{\Delta}_\phi$ to denote $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) \geq \boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\gamma})$ for all $\boldsymbol{\beta},\boldsymbol{\gamma}$, and so $\boldsymbol{\Delta}_\psi \geq 0$ stands for $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) \geq 0, \forall \boldsymbol{\beta},\boldsymbol{\gamma}$. Some basic properties of $\boldsymbol{\Delta}$ are given as follows.

LEMMA 1. *Let ψ and φ be directionally differentiable functions. Then for any $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$, we have the following properties.*

(i) $\boldsymbol{\Delta}_{a\psi+b\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = a\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) + b\boldsymbol{\Delta}_\varphi(\boldsymbol{\beta},\boldsymbol{\gamma})$, $\forall a,b \in \mathbb{R}$.

(ii) If ψ is convex, it is directionally differentiable and $\boldsymbol{\Delta}_\psi \geq 0$; conversely, if ψ is directionally differentiable and $\boldsymbol{\Delta}_\psi \geq 0$ then ψ is convex.

(iii) If $\psi : \mathbb{R}^n \to \mathbb{R}$ is differentiable and $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is continuous and directionally differentiable, then $\boldsymbol{\Delta}_{\psi\circ\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\varphi(\boldsymbol{\beta}),\varphi(\boldsymbol{\gamma})) + \langle \boldsymbol{\Delta}_\varphi(\boldsymbol{\beta},\boldsymbol{\gamma}), \nabla\psi(\varphi(\boldsymbol{\gamma}))\rangle$. Also, if $\psi : \mathbb{R}^n \to \mathbb{R}$ is directionally differentiable and $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ is linear, then $\boldsymbol{\Delta}_{\psi\circ\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\varphi(\boldsymbol{\beta}),\varphi(\boldsymbol{\gamma}))$.

(iv) $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \int_0^1 \big[\delta\psi\big(\boldsymbol{\gamma}+t(\boldsymbol{\beta}-\boldsymbol{\gamma});\boldsymbol{\beta}-\boldsymbol{\gamma}\big) - \delta\psi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})\big]\,\mathrm{d}t$, *provided $\delta\psi(\boldsymbol{\gamma}+t(\boldsymbol{\beta}-\boldsymbol{\gamma});\boldsymbol{\beta}-\boldsymbol{\gamma})$ is integrable over $t \in [0,1]$.*

The properties will be frequently used in the rest of the paper. For instance, for $\psi = \rho\|\cdot\|_2^2/2 - f$, by (i) we can write $\boldsymbol{\Delta}_\psi = \rho\mathbf{D}_2 - \boldsymbol{\Delta}_f$. Sometimes, though f is not necessarily convex, $f + \nu\|\cdot\|_2^2/2$ is so for some $\nu \in \mathbb{R}$, which means $\boldsymbol{\Delta}_f \geq -\nu\mathbf{D}_2$, owing to (ii). For $l(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\alpha})$, commonly encountered in statistical applications, (iii) states that $\boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\alpha},\boldsymbol{X}\boldsymbol{\gamma}+\boldsymbol{\alpha})$. For (iv), the integrability condition is met when the directional derivative restricted to the interval $[\boldsymbol{\beta},\boldsymbol{\gamma}]$ is bounded by a constant (or more generally a Lebesgue integrable function); in particular, if ψ is L-strongly smooth, that is, $\nabla\psi$ exists and is Lipschitz continuous: $\|\nabla\psi(\boldsymbol{\beta}) - \nabla\psi(\boldsymbol{\gamma})\|_* \leq L\|\boldsymbol{\beta}-\boldsymbol{\gamma}\|$ for any $\boldsymbol{\beta},\boldsymbol{\gamma}$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) \leq L\|\boldsymbol{\beta}-\boldsymbol{\gamma}\|^2/2$ and for the Euclidean norm, $\boldsymbol{\Delta}_\psi \leq L\mathbf{D}_2$ results.

Moreover, the GBF operator satisfies some interesting "idempotence" properties under some mild assumptions, which is extremely helpful in studying iterative optimization algorithms.

LEMMA 2. *(i) When ψ is convex, $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) \leq \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma})$, and when ψ is concave, $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) \geq \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma})$ for all $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$.*

(ii) When ψ is directionally differentiable, for all $\boldsymbol{\alpha} = (1-\theta)\boldsymbol{\gamma} + \theta\boldsymbol{\beta}$ with $\theta \notin (0,1)$, $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma})$ and in particular,

$$\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\beta})}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\gamma})}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}). \tag{7}$$

(iii) When $\delta\psi(\cdot;\boldsymbol{\beta}-\boldsymbol{\gamma})$ is bounded in a neighborhood of $\boldsymbol{\alpha}$ and has restricted radial continuity at $\boldsymbol{\alpha}$: $\lim_{\epsilon\to 0+}\delta\psi(\boldsymbol{\alpha}+\epsilon\boldsymbol{h};\boldsymbol{\beta}-\boldsymbol{\gamma}) = \delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\gamma})$ for any $\boldsymbol{h} \in [\boldsymbol{\beta}-\boldsymbol{\alpha},\boldsymbol{\gamma}-\boldsymbol{\alpha}]$, or when $\delta\psi(\boldsymbol{\alpha};\cdot)$ has restricted linearity $\delta\psi(\boldsymbol{\alpha};\boldsymbol{h}) = \langle g(\boldsymbol{\alpha}),\boldsymbol{h}\rangle$ for some g and all $\boldsymbol{h} \in [\boldsymbol{\beta}-\boldsymbol{\alpha},\boldsymbol{\gamma}-\boldsymbol{\alpha}]$, we have

$$\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}). \tag{8}$$

In particular, (8) holds when ψ is differentiable at $\boldsymbol{\alpha}$ or $\delta\psi(\cdot;\boldsymbol{\beta}-\boldsymbol{\gamma})$ is continuous at $\boldsymbol{\alpha}$.

We refer to (ii) as the *weak idempotence* property and (iii) as the *strong idempotence* property. When $\mathbf{\Delta}_\psi$ becomes a legitimate Bregman divergence, (8) can be rephrased into the three-point property $\mathbf{D}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{D}_\psi(\boldsymbol{\beta},\boldsymbol{\alpha}) + \mathbf{D}_\psi(\boldsymbol{\alpha},\boldsymbol{\gamma}) - \langle \boldsymbol{\beta}-\boldsymbol{\alpha}, \nabla\psi(\boldsymbol{\gamma}) - \nabla\psi(\boldsymbol{\alpha})\rangle$ [14]. It is worth mentioning that although from (iii), differentiability can be used to gain strong idempotence, the weak idempotence (7) is often what we need, which always holds under just directional differentiability.

At the end of the subsection, we give some important facts of GBFs for canonical generalized linear models (GLMs) that are widely used in statistics modeling. Here, the response variable $\boldsymbol{y} \in \mathcal{Y}^n \subset \mathbb{R}^n$ has density $p_{\boldsymbol{\eta}}(\cdot) = \exp\{(\langle \cdot, \boldsymbol{\eta}\rangle - b(\boldsymbol{\eta}))/\sigma^2 - c(\cdot,\sigma^2)\}$ with respect to measure ν_0 defined on $\mathcal{Y}^n$ (typically the counting measure or Lebesgue measure), where $\boldsymbol{\eta} \in \mathbb{R}^n$ represents the systematic component of interest, and σ is the scale parameter; see [30]. Since σ is not the parameter of interest, it is more convenient to define the density $\exp\{(\langle \cdot, \boldsymbol{\eta}\rangle - b(\boldsymbol{\eta}))/\sigma^2\}$ (still written as $p_{\boldsymbol{\eta}}(\cdot)$ with a slight abuse of notation) with respect to the base measure $\mathrm{d}\nu = \exp(-c(\cdot,\sigma^2))\,\mathrm{d}\nu_0$. The loss for $\boldsymbol{\eta}$ can be written as

$$l_0(\boldsymbol{\eta};\boldsymbol{y}) = \{-\langle \boldsymbol{y},\boldsymbol{\eta}\rangle + b(\boldsymbol{\eta})\}/\sigma^2. \tag{9}$$

That is, l_0 corresponds to a distribution in the exponential dispersion family with cumulant function $b(\cdot)$, dispersion σ^2 and natural parameter $\boldsymbol{\eta}$. In the Gaussian case, $l_0(\boldsymbol{\eta}) = -\langle \boldsymbol{\eta},\boldsymbol{y}\rangle/\sigma^2 + \|\boldsymbol{y}\|_2^2/(2\sigma^2)$.

Following [62], we define the natural parameter space $\Omega = \mathrm{dom}(b) = \{\boldsymbol{\eta} \in \mathbb{R}^n : b(\boldsymbol{\eta}) < \infty\}$ (always assumed to be nonempty) and the mean parameter space $\mathcal{M} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \boldsymbol{\mu} = \mathbb{E}\boldsymbol{y},$ where $\boldsymbol{y} \sim p$ for some density p defined on $\mathcal{Y}^n$ with respect to $\nu\}$, and call $p_{\boldsymbol{\eta}}$ minimal if $\langle \boldsymbol{a}, \boldsymbol{z}\rangle = c$ for almost every $\boldsymbol{z} \in \mathcal{Y}^n$ with respect to ν implies $\boldsymbol{a} = \boldsymbol{0}$. When Ω is open, $p_{\boldsymbol{\eta}}$ is called regular, and b can be shown to be differentiable to any order and convex, but not necessarily strictly convex; if, in addition, $p_{\boldsymbol{\eta}}$ is minimal, b is strictly convex and the canonical link $g = (\nabla b)^{-1}$ is well-defined on $\mathcal{M}^\circ$. These can all be derived from, say, the propositions in [62].

LEMMA 3. *Assume the exponential dispersion family setup with the associated loss defined in (9). (i) If Ω is an open set or $p_{\boldsymbol{\eta}}$ is regular, then*

$$l_0(\boldsymbol{\eta};\boldsymbol{z}) = \mathbf{\Delta}_b(\boldsymbol{\eta}, \partial b^*(\boldsymbol{z}))/\sigma^2 - b^*(\boldsymbol{z})/\sigma^2 \tag{10}$$

for all $\boldsymbol{\eta} \in \Omega, \boldsymbol{z} \in ri(\mathcal{M})$, where b^* is the Fenchel conjugate of b, and $\partial b^*(\boldsymbol{z})$ can take any subgradient of b^* at $\boldsymbol{z}$. If $p_{\boldsymbol{\eta}}$ is also minimal, $\mathbf{\Delta}_b$ becomes $\mathbf{D}_b$, $\partial b^*(\boldsymbol{z})$ becomes $g(\boldsymbol{z})$ (which is unique), and $ri(\mathcal{M})$ becomes $\mathcal{M}^\circ$. (ii) As long as Ω is open,

$$l_0(\boldsymbol{\eta};\boldsymbol{z}) = \mathbf{\Delta}_{b^*}(\boldsymbol{z}, \nabla b(\boldsymbol{\eta}))/\sigma^2 - b^*(\boldsymbol{z})/\sigma^2 \tag{11}$$

for all $\boldsymbol{\eta} \in \Omega, \boldsymbol{z} \in ri(\mathcal{M})$. If $p_{\boldsymbol{\eta}}$ is also minimal, $\mathbf{\Delta}_{b^*} = \mathbf{D}_{b^*}$ and $ri(\mathcal{M}) = \mathcal{M}^\circ$. (iii) Given any $\boldsymbol{\eta}_1 \in \Omega^\circ$ and $\boldsymbol{\eta}_2 \in \Omega$, the Kullback Leibler (KL) divergence of p_{η_2} from p_{η_1} relates to the GBF of l_0 or b by

$$KL(p_{\boldsymbol{\eta}_1}, p_{\boldsymbol{\eta}_2}) = \mathbf{\Delta}_{l_0}(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1) = \mathbf{\Delta}_b(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1)/\sigma^2. \tag{12}$$

Property (i) shows the importance of GBF in maximum likelihood estimation. A Bregman version of Property (ii) was first described in [3], while our conclusions based on $\mathbf{\Delta}_b, \mathbf{\Delta}_{b^*}$ are more general, as they do *not* require the strict convexity of b or the differentiability of b^*. Consider for instance the multinomial GLM under a symmetric parametrization: for $[y_1,\ldots,y_m] \in \mathcal{Y} = \{y_k \in \{0,1\}, 1 \le k \le m, \sum y_k = 1\}$ $(n = 1)$, $\mathbb{E}y_k \propto \exp(\eta_k)$ or $\mathbb{E}y_k = \exp(\eta_k)/\sum\exp(\eta_k)$ gives $b = \log\sum\exp(\eta_k)$, and thus $b^*(\boldsymbol{\mu})$ takes $\sum \mu_k \log\mu_k$ for

$[\mu_1, \ldots, \mu_m] \in \mathcal{M} = \{[\mu_k] : \sum \mu_k = 1, \mu_k \geq 0\}$ and $+\infty$ otherwise. Clearly, b^* is not differentiable (given any $\boldsymbol{z} \in \text{ri}(\mathcal{M})$, $\partial b^*(\boldsymbol{z}) = \{\log \boldsymbol{z} + t\mathbf{1} : t \in \mathbb{R}\}$), but nicely our two GBF representations still hold. In addition, if the right-hand side of (10) or (11), as a function of $\boldsymbol{z}$, is continuous on $\overline{\mathcal{M}}$, which is the case for Bernoulli, multinomial and Poisson, (i) and (ii) hold for any $\boldsymbol{z} \in \overline{\mathcal{M}}$ from [62, Theorem 3.4].

Property (iii) (notice the exchange of $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ in the generalized Bregman expressions) can be used to formulate and verify model regularity conditions in minimax studies of sparse GLMs, which are of great interest in high-dimensional statistical learning [58]. More concretely, consider a general signal class

$$\mathcal{B}(s^*, M) = \{\boldsymbol{\beta}^* \in \mathbb{R}^p : \|\boldsymbol{\beta}^*\|_0 \leq s^*, \|\boldsymbol{\beta}^*\|_\infty \leq M\}, \tag{13}$$

where $s^* \leq p, 0 \leq M \leq +\infty$. Some applications limit the magnitude of the coefficients β_j via a constraint or a penalty, resulting in a finite M. Let $I(\cdot)$ be any nondecreasing function with $I(0) = 0, I \not\equiv 0$. Some particular examples are $I(t) = t$ and $I(t) = 1_{t \geq c}$. Recall the regular exponential dispersion family with systematic component $\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$ and loss $l(\boldsymbol{\beta}) = l_0(\boldsymbol{\eta})$ defined by (9).

THEOREM 1. *In the regular exponential dispersion family setup (with dom(b) a nonempty open set), assume $p \geq 2, 1 \leq s^* \leq p/2$. Let*

$$P(s^*) = s^* \log(ep/s^*). \tag{14}$$

(i) If

$$\boldsymbol{\Delta}_{l_0}(\mathbf{0}, \boldsymbol{X}\boldsymbol{\beta})\sigma^2 \leq \kappa \mathbf{D}_2(\mathbf{0}, \boldsymbol{\beta}), \quad \forall \boldsymbol{\beta} \in \mathcal{B}(s^*, M) \tag{15}$$

where $\kappa > 0$, there exist positive constants $c, \tilde{c}$, depending on $I(\cdot)$ only, such that

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^* \in \mathcal{B}(s^*, M)} \mathbb{E}\big\{I\big(\mathbf{D}_2(\boldsymbol{\beta}^*, \hat{\boldsymbol{\beta}})/[\tilde{c}\min\{\sigma^2 P(s^*)/\kappa, M^2 s^*\}]\big)\big\} \geq c > 0,$$

where $\hat{\boldsymbol{\beta}}$ denotes any estimator of $\boldsymbol{\beta}^*$.

(ii) If

$$\begin{cases} \underline{\kappa}\mathbf{D}_2(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \leq \mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}_1, \boldsymbol{X}\boldsymbol{\beta}_2) \\ \boldsymbol{\Delta}_{l_0}(\mathbf{0}, \boldsymbol{X}\boldsymbol{\beta}_1)\sigma^2 \leq \overline{\kappa}\mathbf{D}_2(\mathbf{0}, \boldsymbol{\beta}_1), \end{cases} \quad \forall \boldsymbol{\beta}_i \in \mathcal{B}(s^*, M) \tag{16}$$

where $\underline{\kappa}, \overline{\kappa} \geq 0$, then there exist positive constants $c, \tilde{c}$ depending on $I(\cdot)$ only such that

$$\inf_{\hat{\boldsymbol{\beta}}} \sup_{\boldsymbol{\beta}^* \in \mathcal{B}(s^*, M)} \mathbb{E}\big\{I\big(\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{X}\hat{\boldsymbol{\beta}})/[c\min\{(\underline{\kappa}/\overline{\kappa})\sigma^2 P(s^*), \underline{\kappa}M^2 s^*\}]\big)\big\} \geq c > 0.$$

The GBF-form conditions (15), (16) can be viewed as an extension of restricted isometry [11], and are often easy to check using the Hessian. For example, from Lemma 1, we immediately know that if l_0 is L-strongly smooth, (15) is satisfied with $\kappa = L\|X\|_2^2$ even when $M = +\infty$. This is the case for regression and logistic regression, and accordingly, no estimation algorithms can beat the minimax rate $s^* \log(ep/s^*)$ (ignoring trivial factors). The optimal lower bounds provide useful guidance in establishing sharp statistical error upper bounds of Bregman-surrogate algorithms in Section 3.2.

2.2. *Examples of Bregman surrogates.*

EXAMPLE 1. **(Gradient descent and mirror descent)**. Gradient descent is a simple first-order method to minimize a function $f \in \mathcal{C}^1$ which may be nonconvex. Starting with $\boldsymbol{\beta}^{(0)}$, the algorithm proceeds as follows:

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \alpha \nabla f(\boldsymbol{\beta}^{(t)}), \tag{17}$$

where $\alpha > 0$ is a step size parameter. Its rationale can be seen by formulating a Bregman-surrogate algorithm using $\boldsymbol{\Delta}_\psi = \rho \mathbf{D}_2 - \boldsymbol{\Delta}_f$:

$$\boldsymbol{\beta}^{(t+1)} = \arg\min_{\boldsymbol{\beta}} \; g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}) + (\rho \mathbf{D}_2 - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) \tag{18a}$$

$$= \boldsymbol{\beta}^{(t)} - \frac{1}{\rho} \nabla f(\boldsymbol{\beta}^{(t)}), \tag{18b}$$

where $f(\cdot) - \boldsymbol{\Delta}_f(\cdot, \boldsymbol{\beta}^{(t)})$ gives a linear approximation of f and $1/\rho$ amounts to the step size. We call ρ the inverse step size parameter. (The generalized Bregman surrogate in (18a) extends the class of algorithms to a directionally differentiable f, with the update given by $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + (0 \vee -\delta f(\boldsymbol{\beta}^{(t)}; \boldsymbol{h}^\circ))\boldsymbol{h}^\circ/\rho$ and $\boldsymbol{h}^\circ \in \arg\max_{\|\boldsymbol{h}\|_2=1} [\delta f(\boldsymbol{\beta}^{(t)}; \boldsymbol{h})]_-$, where $[]_-$ denotes the negative part ($t_- = (|t| - t)/2$).)

More generally, we can use a strictly convex $\varphi \in \mathcal{C}^1$ to construct

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}) + (\rho \mathbf{D}_\varphi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}), \tag{19}$$

Minimizing (19) with respect to $\boldsymbol{\beta}$ gives the renowned mirror descent [38]: $\boldsymbol{\beta}^{(t+1)} = (\nabla\varphi)^{-1}(\nabla\varphi(\boldsymbol{\beta}^{(t)}) - \nabla f(\boldsymbol{\beta}^{(t)})/\rho)$, where $(\nabla\varphi)^{-1}$ is the inverse of $\nabla\varphi$. Mirror descent is widely used in convex programming, but this work does *not* restrict f to be convex.

EXAMPLE 2. **(Iterative thresholding)**. Sparsity-inducing penalties are widely used in high-dimensional problems; see, for example, ℓ_0, ℓ_1 [56], bridge penalties [22], SCAD [19], capped-ℓ_1 [66] and MCP [65]. There is a universal connection between thresholding rules and penalty functions [48], and the mapping from penalties to thresholdings is many-to-one. This makes it possible to apply an iterative thresholding algorithm to solve a general penalized problem of the form $\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + \sum_j P(\varrho\beta_j; \lambda)$ [8, 47]:

$$\boldsymbol{\beta}^{(t+1)} = \Theta(\varrho\boldsymbol{\beta}^{(t)} - \nabla l(\boldsymbol{\beta}^{(t)})/\varrho; \lambda)/\varrho, \tag{20}$$

where Θ is a thresholding function inducing P, and $\varrho > 0$ is an algorithm parameter for the sake of scaling and convergence control. This class of iterative algorithms is called the *Thresholding-based Iterative Selection Procedures* (TISP) in [47] and is scalable in computation. For the rigorous definition of Θ and the Θ-P coupling formula, see Section 3.1 for detail. Some examples of Θ include: (i) soft-thresholding $\Theta_S(t; \lambda) = \mathrm{sgn}(t)(|t| - \lambda)1_{|t|>\lambda}$, which induces the ℓ_1 penalty, (ii) hard-thresholding $\Theta_H(t; \lambda) = t1_{|t|>\lambda}$, which is associated with (infinitely) many penalties, with the capped-ℓ_1 penalty, (55), and the discrete ℓ_0 penalty as particular instances. The nonconvex SCAD and MCP penalties also have their corresponding thresholding rules. In this sense, thresholdings extend proximity operators. One can regard (20) as an outcome of minimizing the following Bregman surrogate

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = l(\boldsymbol{\beta}) + \sum P(\varrho\beta_j; \lambda) + (\varrho^2 \mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}). \tag{21}$$

Here, we linearize l only, as $\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)})$ has (20) as its globally optimal solution. Interestingly, the set of fixed points under the g-mapping enjoys provable guarantees that may *not*

hold for the set of local minimizers to the original objective (Section 3.2.1). This is particularly the case when Θ has discontinuities and $P(t;\lambda)$ is given by $P_\Theta(t;\lambda)+q(t;\lambda)$, where P_Θ is defined by (48) and q is a function satisfying $q(t;\lambda)\geq 0$ for all $t\in\mathbb{R}$ and $q(t;\lambda)=0$ if $t=\Theta(s;\lambda)$ for some $s\in\mathbb{R}$ [49].

A closely related *iterative quantile-thresholding* procedure [48, 52] proceeds by $\boldsymbol{\beta}^{(t+1)}=\Theta^{\#}(\boldsymbol{\beta}^{(t)}-\nabla l(\boldsymbol{\beta}^{(t)})/\varrho^2;q)$ for the sake of feature screening: $\min l(\boldsymbol{\beta})$ s.t. $\|\boldsymbol{\beta}\|_0\leq q$, and uses a similar surrogate $g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)})=l(\boldsymbol{\beta})+(\varrho^2\mathbf{D}_2-\boldsymbol{\Delta}_l)(\boldsymbol{\beta},\boldsymbol{\beta}^{(t)})$. Here, the quantile thresholding $\Theta^{\#}(\boldsymbol{\alpha};q)$, as an outcome of $\min g(\boldsymbol{\beta};\boldsymbol{\beta}^{-})$, keeps the top q elements of α_j after ordering them in magnitude, $|\alpha_{(1)}|\geq\cdots\geq|\alpha_{(p)}|$, and zero out the rest. To avoid ambiguity, we assume no ties occur in performing $\Theta^{\#}(\boldsymbol{\alpha};q)$ throughout the paper, that is, $|\alpha_{(q)}|>|\alpha_{(q+1)}|$.

EXAMPLE 3. **(Nonnegative matrix factorization)**. Nonnegative Matrix Factorization (NMF) [34] provides an effective tool for feature extraction and finds widespread applications in computer vision, text mining and many other areas. NMF approximates a nonnegative data matrix $\boldsymbol{X}\in\mathbb{R}_+^{n\times p}$ by the product of two nonnegative low-rank matrices $\boldsymbol{W}\in\mathbb{R}_+^{n\times r}$ and $\boldsymbol{H}\in\mathbb{R}_+^{r\times p}$. The KL divergence is often used to make a cost function, that is, $\min_{\boldsymbol{W}\in\mathbb{R}_+^{n\times r},\boldsymbol{H}\in\mathbb{R}_+^{r\times p}}\mathrm{KL}(\boldsymbol{X},\boldsymbol{W}\boldsymbol{H}):=\sum_{i,j}[X_{ij}\log(X_{ij}/(\boldsymbol{W}\boldsymbol{H})_{ij})-X_{ij}+(\boldsymbol{W}\boldsymbol{H})_{ij}]$, which gives a nonconvex optimization problem. The following *multiplicative* update rule (MUR) shows good scalability in big data applications [15]:

$$H_{kj}^{(t+1)}=H_{kj}^{(t)}\exp\Big[-\frac{1}{\rho}\sum_i\Big(W_{ik}-\frac{W_{ik}X_{ij}}{(\boldsymbol{W}\boldsymbol{H}^{(t)})_{ij}}\Big)\Big], \tag{22}$$

$$W_{ik}^{(t+1)}=W_{ik}^{(t)}\exp\Big[-\frac{1}{\rho}\sum_j\Big(H_{kj}-\frac{H_{kj}X_{ij}}{(\boldsymbol{W}^{(t)}\boldsymbol{H})_{ij}}\Big)\Big]. \tag{23}$$

The update formulas can be explained from a Bregman surrogate perspective. Since the problem is symmetric in $\boldsymbol{W}$ and $\boldsymbol{H}$, $\boldsymbol{\Delta}_{\mathrm{KL}}(\boldsymbol{X},\boldsymbol{W}\boldsymbol{H})=\boldsymbol{\Delta}_{\mathrm{KL}}(\boldsymbol{X}^\top,\boldsymbol{H}^\top\boldsymbol{W}^\top)$, we take (22) for instance to illustrate the point. Noticing that the criterion is separable in the column vectors of $\boldsymbol{H}$, it suffices to look at $\min_{\boldsymbol{h}\in\mathbb{R}_+^r}f(\boldsymbol{h})=\mathrm{KL}(\boldsymbol{x},\boldsymbol{W}\boldsymbol{h})=\sum_i[x_i\log(x_i/(\boldsymbol{W}\boldsymbol{h})_i)-x_i+(\boldsymbol{W}\boldsymbol{h})_i]$, where $\boldsymbol{x}$ can be any column of $\boldsymbol{X}$. Then it is easy to verify that the following Bregman surrogate,

$$g(\boldsymbol{h};\boldsymbol{h}^{(t)})=f(\boldsymbol{h})+(\rho\mathbf{D}_\varphi-\mathbf{D}_f)(\boldsymbol{h},\boldsymbol{h}^{(t)}),\ \varphi(\boldsymbol{h})=\sum(h_i\log h_i-h_i), \tag{24}$$

leads to the multiplicative update formulas.

EXAMPLE 4. **(DC programming)**. DC programming [55] is capable of tackling a large class of nonsmooth nonconvex optimization problems; see, for example, [23, 43]. A "difference of convex" (DC) function f is defined by $f(\boldsymbol{\beta})=d_1(\boldsymbol{\beta})-d_2(\boldsymbol{\beta})$, where d_1 and d_2 are both closed convex functions. To minimize $f(\boldsymbol{\beta})$, a standard DC algorithm generates two sequences $\{\boldsymbol{\beta}^{(t)}\}$ and $\{\boldsymbol{\gamma}^{(t)}\}$ that obey

$$\boldsymbol{\gamma}^{(t)}\in\partial d_2(\boldsymbol{\beta}^{(t)}),\ \boldsymbol{\beta}^{(t+1)}\in\partial d_1^*(\boldsymbol{\gamma}^{(t)}), \tag{25}$$

where $\partial d(\boldsymbol{\beta})$ is the subdifferential of $d(\cdot)$ at $\boldsymbol{\beta}$, and $d_1^*(\cdot)$ is the Fenchel conjugate of $d_1(\cdot)$. (As before, d_1,d_2 are assumed to be real-valued functions defined on $\mathbb{R}^p$, so the sequences are well-defined and finite.) This elegant algorithm does not involve any line search and guarantees global convergence given any initial point. Many popular nonconvex algorithms can be derived from (25) [2].

Focusing on the $\boldsymbol{\beta}$-update, we know that $\boldsymbol{\beta}^{(t+1)}$ must be a solution to $\min_{\boldsymbol{\beta}}d_1(\boldsymbol{\beta})-\langle\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)}\rangle$ or $\min_{\boldsymbol{\beta}}d_1(\boldsymbol{\beta})-\langle\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}\rangle$. Due to the convexity of d_2, $\langle\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}\rangle\leq$

$\sup_{\boldsymbol{\gamma}\in\partial d_2(\boldsymbol{\beta}^{(t)})}\langle\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}\rangle = \delta d_2(\boldsymbol{\beta}^{(t)};\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})$ for all $\boldsymbol{\gamma}^{(t)}\in\partial d_2(\boldsymbol{\beta}^{(t)}), \boldsymbol{\beta}\in\mathbb{R}^p$. Thus $\min_{\boldsymbol{\beta}} d_1(\boldsymbol{\beta}) - \langle\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}\rangle$ should be no lower than $\min_{\boldsymbol{\beta}} d_1(\boldsymbol{\beta}) - \delta d_2(\boldsymbol{\beta}^{(t)};\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})$. Choosing $\boldsymbol{\beta}^{(t+1)}\in\arg\min d_1(\boldsymbol{\beta}) - \delta d_2(\boldsymbol{\beta}^{(t)};\boldsymbol{\beta}-\boldsymbol{\beta}^{(t)})$ and $\boldsymbol{\gamma}^{(t)} = \delta d_2(\boldsymbol{\beta}^{(t)};\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)})(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)})/\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_2^2$ ensures (25), which simply amounts to using a Bregman surrogate

$$g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{d_2}(\boldsymbol{\beta},\boldsymbol{\beta}^{(t)}). \tag{26}$$

For the $\boldsymbol{\gamma}$-updates, a Bregman surrogate $g(\boldsymbol{\gamma};\boldsymbol{\gamma}^{(t)}) = (d_2^*-d_1^*)(\boldsymbol{\gamma}) + \boldsymbol{\Delta}_{d_1^*}(\boldsymbol{\gamma},\boldsymbol{\gamma}^{(t)})$ can be similarly constructed.

EXAMPLE 5. **(Local linear approximation)**. Zou and Li [67] proposed an effective local linear approximation (LLA) technique to minimize penalized negative log-likelihoods. In their paper, the loss function is assumed to be convex and smooth, and the penalty is concave on $\mathbb{R}_+$. We give a new characterization of LLA by use of a Bregman surrogate.

Let l be a directionally differentiable loss function but not necessarily continuously differentiable, and P be a function that is concave and differentiable over $(0,+\infty)$, and satisfies $P(t) = P(-t)$ for any $t\in\mathbb{R}$, $P(0) = 0$. Consider the problem $\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + \sum_j P(\beta_j)$. Using the generalized Bregman notation $\boldsymbol{\Delta}_{\|\cdot\|_1}(\boldsymbol{\beta},\boldsymbol{\gamma})$, or $\boldsymbol{\Delta}_1(\boldsymbol{\beta},\boldsymbol{\gamma})$ for short, define

$$g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)}) = l(\boldsymbol{\beta}) + \sum P(\beta_j) + \sum\left[\alpha_j\boldsymbol{\Delta}_1(\beta_j,\beta_j^{(t)}) - \boldsymbol{\Delta}_P(\beta_j,\beta_j^{(t)})\right]. \tag{27}$$

In contrast to (21), (27) linearizes P instead of l. Simple calculation shows

$$\boldsymbol{\Delta}_1(\beta_j,\beta_j^{(t)}) = \begin{cases} |\beta_j| - \operatorname{sgn}(\beta_j^{(t)})\beta_j, & \beta_j^{(t)}\neq 0 \\ 0, & \beta_j^{(t)} = 0, \end{cases} \tag{28}$$

$$\boldsymbol{\Delta}_P(\beta_j,\beta_j^{(t)}) = \begin{cases} P(\beta_j) - P(\beta_j^{(t)}) - P'(\beta_j^{(t)})(\beta_j-\beta_j^{(t)}), & \beta_j^{(t)}\neq 0 \\ P(\beta_j) - P'_+(0)|\beta_j|, & \beta_j^{(t)} = 0, \end{cases} \tag{29}$$

where $\operatorname{sgn}(\cdot)$ is the sign function and $P'_+(\beta)$ denotes the right derivative of $P(\cdot)$ at β. Interestingly, with $\alpha_j = |P'_+(\beta_j^{(t)})|$, the $\boldsymbol{\Delta}_1$-based surrogate (27) can be shown to be

$$l(\boldsymbol{\beta}) + \sum_j\left[P(|\beta_j^{(t)}|) + P'_+(|\beta_j^{(t)}|)(|\beta_j| - |\beta_j^{(t)}|)\right],$$

which is exactly the surrogate constructed by Zou and Li. To the best of our knowledge, the generalized Bregman formulation is new.

LLA requires solving a weighted lasso problem at each step. We can further linearize l as in Example 2 to improve its scalability. LLA is popular among statisticians, but to our knowledge, there is a lack of *global* convergence-rate studies in large-p applications. We will see that reformulating LLA from the generalized Bregman surrogate perspective leads to a convenient choice of the convergence measure in analyzing the algorithm.

EXAMPLE 6. **(Sigmoidal regression)**. We use the univariate-response sigmoidal regression to illustrate this type of nonconvex problems that is commonly seen in artificial neural networks. The formulation carries over to multilayered networks and recurrent networks [51].

Let $\boldsymbol{X} = [\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]^\top\in\mathbb{R}^{n\times p}$ be the data matrix, and $\boldsymbol{y} = [y_1,\cdots,y_n]^\top$ be the response vector. Define $\pi(\nu) = e^\nu/(1+e^\nu)$; if ν is replaced by a vector, π is defined componentwise. The sigmoidal regression solves

$$\min_{\boldsymbol{\beta}}\ f(\boldsymbol{\beta}) = \frac{1}{2}\sum_{i=1}^n(y_i - \pi(\boldsymbol{x}_i^\top\boldsymbol{\beta}))^2. \tag{30}$$

Then $\nabla^2 f(\boldsymbol{\beta}) = \sum_{i=1}^n [(-2\mu_i^3 + 3\mu_i^2 - \mu_i)y_i + (3\mu_i^4 - 5\mu_i^3 + 2\mu_i^2)]\boldsymbol{x}_i\boldsymbol{x}_i^\top$, where $\mu_i = \pi(\boldsymbol{x}_i^\top\boldsymbol{\beta})$. Because $\mu_i \in [0,1]$, we get $\nabla^2 f(\boldsymbol{\beta}) \preceq \boldsymbol{X}^\top \mathrm{diag}\{|0.1y_i| + 0.08\}_{i=1}^n \boldsymbol{X}$, which motivates a Bregman surrogate

$$g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)}) = f(\boldsymbol{\beta}) + \mathbf{D}_{\psi-f}(\boldsymbol{\beta},\boldsymbol{\beta}^{(t)}),\ \psi(\boldsymbol{\beta}) = \frac{1}{2}\boldsymbol{\beta}^\top\boldsymbol{X}^\top\mathrm{diag}\{|0.1y_i| + 0.08\}\boldsymbol{X}\boldsymbol{\beta}.$$

Solving $\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)})$ yields $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \boldsymbol{B}^{-1}\boldsymbol{X}^\top(\boldsymbol{u}^{(t)} - \boldsymbol{u}^{(t)} \circ \boldsymbol{u}^{(t)}) \circ (\boldsymbol{y} - \boldsymbol{u}^{(t)})$, where $\boldsymbol{B} = \boldsymbol{X}^\top \mathrm{diag}\{|0.1y_i| + 0.08\}_{i=1}^n \boldsymbol{X}$, $\boldsymbol{u}^{(t)} = \pi(\boldsymbol{X}^\top\boldsymbol{\beta}^{(t)})$ and $\circ$ denotes the Hadamard product. This type of surrogate functions is closely related to proximal Newton-type methods [46] and signomial programming [33].

3. Bregman-surrogate algorithm analysis. Motivated by the examples in Section 2, we study a generalized Bregman-surrogate algorithm family for solving $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta})$, with the sequence of iterates defined by

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}}\ g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)}) := f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\beta}^{(t)}),\ t \geq 0 \tag{31}$$

The objective function f and the auxiliary function ψ are assumed to be directionally differentiable but need not be smooth or convex. ψ has flexible options as seen from the previous examples.

Equation (31) does not necessarily give an MM procedure, as the majorization condition $g(\boldsymbol{\beta};\boldsymbol{\beta}^-) \geq f(\boldsymbol{\beta})$ may not hold. But we have the following zeroth-order *and* first-order degeneracies when $\boldsymbol{\beta}^- = \boldsymbol{\beta}$, which provides rationality of investigating the accuracy of *fixed points* under the g-mapping (31).

LEMMA 4. *Let $g(\boldsymbol{\beta};\boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\beta}^-)$ with f and ψ directionally differentiable. Then (i) $g(\boldsymbol{\beta};\boldsymbol{\beta}) = f(\boldsymbol{\beta})$, and (ii) $\delta g(\boldsymbol{\beta};\boldsymbol{\beta}^-,\boldsymbol{h})|_{\boldsymbol{\beta}^-=\boldsymbol{\beta}} = \delta f(\boldsymbol{\beta};\boldsymbol{h}), \forall \boldsymbol{\beta},\boldsymbol{h}$, where $\delta g(\boldsymbol{\beta};\boldsymbol{\beta}^-,\boldsymbol{h})$ is the directional derivative of $g(\,\cdot\,;\boldsymbol{\beta}^-)$ at $\boldsymbol{\beta}$ with increment $\boldsymbol{h}$.*

The lemma relates the set of fixed points of the algorithm mapping,

$$\{\boldsymbol{\beta} : \boldsymbol{\beta} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta};\boldsymbol{\beta}^-)|_{\boldsymbol{\beta}^-=\boldsymbol{\beta}}\}, \tag{32}$$

which we will call the fixed points of g for short, to the set of directional stationary points of f (under directional differentiability),

$$\{\boldsymbol{\beta} : \delta f(\boldsymbol{\beta};\boldsymbol{h}) \geq 0 \text{ for any admissible } \boldsymbol{h}\}, \tag{33}$$

which becomes the set of stationary points when $f \in \mathcal{C}^1$. The link is general for any generalized Bregman surrogate in (31) *regardless* of the specific form of ψ. An important implication is that in studying convergence it is legitimate to measure how $\boldsymbol{\beta}^{(t+1)}$ and $\boldsymbol{\beta}^{(t)}$ differ, as widely used in practice. Later we will see that it is indeed possible to provide provable guarantees for the fixed points of this type of surrogates. In contrast, a general MM algorithm does not always have the first-order degeneracy and so attaining $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$ does not necessarily ensure a good-quality solution, especially in nonconvex scenarios.

3.1. *Computational accuracy.* We first study the optimization error of (31), then turn to its statistical error in Section 3.2. This subsection aims to derive universal rates of convergence under no regularity conditions.

• *General setting.* In this part, the objective $f(\boldsymbol{\beta})$ does not have any known structure. To better connect with some conventional results in convex optimization, we first present two propositions for (31) on the function-value convergence and iterate convergence. While the resultant rates are encouraging, the error bounds are most informative under certain smoothness and convexity assumptions. This suggests the necessity of choosing a proper convergence measure in order to avoid stringent or awkward technical conditions in nonconvex optimization.

PROPOSITION 1. *Given an arbitrary initial point $\boldsymbol{\beta}^{(0)}$, let $\boldsymbol{\beta}^{(t)}$ be the sequence generated according to* (31) *where ψ is differentiable. Then*

$$\operatorname*{avg}_{0\le t\le T} f(\boldsymbol{\beta}^{(t+1)}) - f(\bar{\boldsymbol{\beta}}) \le \frac{1}{T+1}[\boldsymbol{\Delta}_\psi(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^{(0)}) - \boldsymbol{\Delta}_\psi(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^{(T+1)})] \tag{34}$$

for any $\bar{\boldsymbol{\beta}}$ satisfying

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + \boldsymbol{\Delta}_f(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^{(t+1)}) \ge 0,\ 0 \le t \le T. \tag{35}$$

Here, $\operatorname{avg}_{0\le t\le T} f(\boldsymbol{\beta}^{(t+1)})$ *denotes the average of* $f(\boldsymbol{\beta}^{(1)}), \ldots, f(\boldsymbol{\beta}^{(T+1)})$.

In particular, if both f and ψ are convex, then $f(\boldsymbol{\beta}^{(t)})$ is nonincreasing and

$$f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) \le \frac{\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)})}{T+1},\ \forall \boldsymbol{\beta}. \tag{36}$$

Equation (34) shows a convergence rate of $\mathcal{O}(1/T)$ under (35) that amounts to step size control. For example, for $\boldsymbol{\Delta}_\psi = \rho \mathbf{D}_\varphi - \boldsymbol{\Delta}_f$ in mirror descent, (35) shows that ρ should be sufficiently large, which in turns gives a small stepsize $1/\rho$:

$$\rho \ge (\boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) - \boldsymbol{\Delta}_f(\bar{\boldsymbol{\beta}}, \boldsymbol{\beta}^{(t+1)}))/\mathbf{D}_\varphi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}),$$

or $\rho \ge \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)})/\mathbf{D}_\varphi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)})$ when f is convex. In nonconvex scenarios, the condition may be hard to verify, but one has reason to believe that with a properly small step size, a generalized Bregman-surrogate algorithm should not be much slower than gradient descent.

Actually, a faster rate of convergence may be obtained under some GBF comparison conditions, (37) and (39) below, which can be viewed as substitutes for conventional strong convexity in a more general sense. (The corresponding geometric decay of the errors is motivating in high dimensional statistical learning, in light of the "restricted" strongly convexity often possessed by such a type of problems [36].)

PROPOSITION 2. *Consider the iterative algorithm defined by* (31) *starting at an arbitrary point $\boldsymbol{\beta}^{(0)}$ with ψ differentiable, and let $\boldsymbol{\beta}^o$ be a minimizer of $f(\boldsymbol{\beta})$. (i) If for some $\kappa > 1$, $\boldsymbol{\Delta}_\phi = \boldsymbol{\Delta}_\psi + \boldsymbol{\Delta}_f$ satisfies*

$$\bar{\boldsymbol{\Delta}}_\phi \ge \frac{\kappa}{\kappa - 1}\boldsymbol{\Delta}_\psi, \tag{37}$$

then for any $T \ge 0$, we have

$$\bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(T+1)}) \le \left(\frac{\kappa-1}{\kappa+1}\right)^{T+1} \bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(0)}) - \frac{\kappa}{2} \min_{0\le t\le T} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}). \tag{38}$$

(ii) Alternatively, if

$$2\bar{\boldsymbol{\Delta}}_f \ge \varepsilon \boldsymbol{\Delta}_\psi \tag{39}$$

for some $\varepsilon > 0$, then

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(T+1)}) \le \Big(\frac{1}{1+\varepsilon}\Big)^{T+1} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(0)}) - \frac{1}{\varepsilon} \min_{0\le t\le T} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \tag{40}$$

for any $T \ge 0$.

REMARK 1. *We give an illustration of (i) and (ii) to compare their assumptions and conclusions. In gradient descent with $\boldsymbol{\Delta}_\phi = \rho \mathbf{D}_2$, (37) becomes $\rho \mathbf{D}_2 \ge (\rho \mathbf{D}_2 - \boldsymbol{\Delta}_f)\kappa/(\kappa - 1)$ or $\boldsymbol{\Delta}_f \ge (\rho/\kappa)\mathbf{D}_2$ and when f is μ-strongly convex and ρ-strongly smooth, $\kappa = \rho/\mu$. Then (38) reads*

$$\mathbf{D}_2(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(T+1)}) \le \Big(\frac{\rho-\mu}{\rho+\mu}\Big)^{T+1} \mathbf{D}_2(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(0)}). \tag{41}$$

The $\mathbf{D}_2$-form bound is classical for problems with strong convexity; see, for example, Theorem 2.1.15 in [41]. Yet it is worth mentioning that our Bregman comparison conditions do not require ψ to be strongly *convex to attain the linear rate. (40) gives a linear convergence result, too, in terms of yet another measure. In the same setup, (39) holds for $\varepsilon : \varepsilon\rho/(2+\varepsilon) = \mu$ and similarly*

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(T+1)}) \le \Big(\frac{\rho-\mu}{\rho+\mu}\Big)^{T+1} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(0)}). \tag{42}$$

A careful examination of the proof in Section A.8 shows that (39) is applied once, while (37) is applied twice on both sides of (A.13), and so (ii) appears less technically demanding. Picking a suitable error function can assist analysis and relax regularity assumptions. The same $\boldsymbol{\Delta}_\psi$ will be used in studying the statistical error convergence in Theorem 5.

Instead of naively comparing $f(\boldsymbol{\beta}^{(t)})$ with f^o, or $\boldsymbol{\beta}^{(t)}$ with $\boldsymbol{\beta}^o$, which may be unattainable or nonunique in nonconvex optimization, one can measure the algorithm convergence in a wiser manner. Ben-Tal and Nemirovski [5] pointed out that with an inappropriate measure of discrepancy, the convergence rate of gradient descent for minimizing a nonconvex objective can be arbitrarily slow, and a common choice is to bound

$$\min_{t\le T} \|\nabla f(\boldsymbol{\beta}^{(t)})\|^2. \tag{43}$$

This is reasonable since when $\nabla f(\boldsymbol{\beta}^{(t)}) = 0$, gradient descent stops iterating and delivers a stationary point. (43) can be rewritten as ρ^2 times

$$\min_{t\le T} \mathbf{D}_2(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \tag{44}$$

as $\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\beta}^{(t)} = -\nabla f(\boldsymbol{\beta}^{(t)})/\rho$. The idea of checking stationarity by the difference between two successive iterates generalizes, thanks to Lemma 4, and eventually leads to an error bound that can get rid of condition (35).

THEOREM 2. *Any generalized Bregman surrogate algorithm defined by (31) satisfies the following bound for all $T \ge 1$,*

$$\operatorname*{avg}_{0\le t\le T} (2\bar{\boldsymbol{\Delta}}_\psi + \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le \frac{1}{T+1}\big[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})\big]. \tag{45}$$

(45) obtains the same rate of convergence as Proposition 1, but is *free* of any conditions other than directional differentiability, because only the weak idempotence is needed to derive

the bound. A proper stepsize control can often make the GBF error nonnegative (e.g., (50)). But even when $\boldsymbol{\beta}^{(t)}$ diverges, (45) still applies.

Notice the factor '2' proceeding the symmetrized Bregman $\bar{\boldsymbol{\Delta}}_\psi$ on the left-hand side of (45). This gives a relaxed stepsize control than MM. We use mirror descent $\boldsymbol{\Delta}_\psi = \rho \mathbf{D}_\varphi - \boldsymbol{\Delta}_f$ to exemplify the point without requiring f to be convex, cf. Example 1.

COROLLARY 1. *In the mirror descent setup with a possibly nonconvex objective, suppose that $\boldsymbol{\Delta}_f \le L\bar{\mathbf{D}}_\varphi$ for some $L > 0$, $\inf_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \ge 0$, and the inverse stepsize parameter ρ is taken such that $\rho > L/2$. Then any accumulation point of $\boldsymbol{\beta}^{(t)}$ is a fixed point of g and*

$$\operatorname*{avg}_{0\le t\le T} \bar{\mathbf{D}}_\varphi(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le \frac{f(\boldsymbol{\beta}^{(0)})}{(T+1)(2\rho - L)}. \tag{46}$$

Hence in the special case of gradient descent, (46) recovers $\min_{0\le t\le T} \|\nabla f(\boldsymbol{\beta}^{(t)})\|_2^2 = \mathcal{O}(1/T)$ [5] when $\rho > L/2$. In comparison, MM algorithms always require $\boldsymbol{\Delta}_\psi \ge 0$, or $\rho \ge L$. A smaller value of ρ means a larger step size with which the algorithm converges faster.

• *Composite setting.* High-dimensional statistical learning often has an additive objective $f(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}; \lambda)$, where $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is the predictor or feature matrix, $l_0(\cdot)$ is the loss defined on $\boldsymbol{X}\boldsymbol{\beta}$ (and so $l(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta})$), $P(\cdot;\lambda)$ is a sparsity-inducing regularizer and ϱ is a controllable parameter, typically taking $\|\boldsymbol{X}\|_2$ to match the scale. Unless otherwise mentioned, $P(\boldsymbol{\beta};\lambda)$ denotes $\sum_j P(\beta_j;\lambda)$ with a little abuse of notation.

Such a composite setup is widely assumed in convex optimization [57, 18]. But among the abundant choices of l_0 and P in the literature, many of them are nonconvex. The good news is that the main theorem proved in the previous subsection adapts to the composite setting and we give some results for iterative thresholding and LLA as an illustration (cf. Examples 2, 5).

Iterative thresholding. Many popularly used penalty functions are associated with thresholdings rigorously defined as follows.

DEFINITION 3 (**Thresholding function**). *A threshold function is a real-valued function $\Theta(t;\lambda)$ defined for $-\infty < t < \infty$ and $0 \le \lambda < \infty$ such that (i) $\Theta(-t;\lambda) = -\Theta(t;\lambda)$; (ii) $\Theta(t;\lambda) \le \Theta(t';\lambda)$ for $t \le t'$; (iii) $\lim_{t\to\infty} \Theta(t;\lambda) = \infty$; (iv) $0 \le \Theta(t;\lambda) \le t$ for $0 \le t < \infty$.*

Given Θ, a critical concavity number $\mathcal{L}_\Theta \le 1$ can be introduced such that $\mathrm{d}\Theta^{-1}(u;\lambda)\,\mathrm{d}u \ge 1 - \mathcal{L}_\Theta$ for almost every $u \ge 0$, or

$$\mathcal{L}_\Theta = 1 - \operatorname{ess\,inf}\{\mathrm{d}\Theta^{-1}(u;\lambda)/\mathrm{d}u : u \ge 0\}, \tag{47}$$

with ess inf the essential infimum and $\Theta^{-1}(u;\lambda) := \sup\{t : \Theta(t;\lambda) \le u\}, \forall u > 0$. For the widely used soft-thresholding $\Theta_S(t;\lambda) = \operatorname{sgn}(t)(|t| - \lambda)1_{|t|>\lambda}$ and hard-thresholding $\Theta_H(t;\lambda) = t1_{|t|>\lambda}$, $\mathcal{L}_\Theta$ equals 0 and 1, respectively. In fact, when $\mathcal{L}_\Theta > 0$, the penalty induced by Θ via (48) is nonconvex, and $\mathcal{L}_\Theta$ gives a concavity measure of it according to Lemma A.3. The Bregman surrogate characterization of iterative thresholding in (21) yields a general conclusion for any Θ in possibly high dimensions.

PROPOSITION 3. *Given any thresholding Θ and directionally differentiable $l(\cdot)$, consider the iterative thresholding procedure (20): $\boldsymbol{\beta}^{(t+1)} = \Theta(\varrho\boldsymbol{\beta}^{(t)} - \nabla l(\boldsymbol{\beta}^{(t)})/\varrho; \lambda)/\varrho$ with $\varrho > 0$. Construct*

$$P_\Theta(t;\lambda) = \int_0^{|t|} (\Theta^{-1}(u;\lambda) - u)\,\mathrm{d}u,\ \forall t \in \mathbb{R}, \tag{48}$$

and define $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda)$, $g(\boldsymbol{\beta}, \boldsymbol{\beta}^-) = l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) + (\varrho^2\mathbf{D}_2 - \boldsymbol{\Delta}_l)(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$. *Then* $\boldsymbol{\beta}^{(t)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t-1)})$ *and for all* $T \geq 1$

$$\operatorname*{avg}_{0\leq t\leq T} (\varrho^2(2 - \mathcal{L}_\Theta)\mathbf{D}_2 - \hat{\boldsymbol{\Delta}}_l)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \leq \frac{1}{T+1}\big[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})\big]. \tag{49}$$

When the loss satisfies $\boldsymbol{\Delta}_l \leq L\mathbf{D}_2$, a reasonable choice of ϱ is

$$\varrho^2 > L/(2 - \mathcal{L}_\Theta). \tag{50}$$

So when $\mathcal{L}_\Theta > 0$, the step size upper bound will be smaller than that as $\mathcal{L}_\Theta = 0$. This is often the price to pay for nonconvex optimization. On the other hand, (49) still ensures the universal rate of convergence of $\mathcal{O}(1/T)$, in spite of the high dimensionality and nonconvexity.

Local linear approximation. Next, we study the computational convergence of LLA for solving the penalized estimation problem $\min f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta})$, assuming l is directionally differentiable, $P(0) = 0$, $P'_+(0) < +\infty$, $P(t) = P(-t) \geq 0$ and $P(t)$ is differentiable for any $t > 0$. Recall its Bregman form surrogate

$$g^{(t)}_{\mathrm{LLA}}(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1 - P(\cdot)}(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^{(t)}), \tag{51}$$

where $\boldsymbol{\alpha}^{(t)} = [\alpha^{(t)}_j]$ with $\alpha^{(t)}_j = |P'_+(\beta^{(t)}_j)|, 1 \leq j \leq p$. We abbreviate $\boldsymbol{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1 - P(\cdot)}$ to $\boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}$, which does not satisfy strong idempotence. By combining $\bar{\boldsymbol{\Delta}}^{(t)}_{\mathrm{LLA}}$ and $\boldsymbol{\Delta}_f$ to evaluate LLA's optimization error, we obtain a convergence result without any additional assumptions.

PROPOSITION 4. *Given any starting point* $\boldsymbol{\beta}^{(0)}$*, the LLA iterates satisfy the following bound for all* $T \geq 1$*:*

$$\operatorname*{avg}_{0\leq t\leq T} [2\bar{\boldsymbol{\Delta}}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t)}, \varrho\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)})] \leq \frac{1}{T+1}[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})].$$

Ignoring the cost difference per iteration, the convergence rate of LLA is no slower than that of gradient descent. If l is a negative log-likelihood function associated with a log-concave density and P is concave on $\mathbb{R}_+$, as assumed in [67], $2\bar{\boldsymbol{\Delta}}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}') = \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + \boldsymbol{\Delta}_{-P}(\varrho\boldsymbol{\beta}', \varrho\boldsymbol{\beta}) + 2\sum_j \alpha^{(t)}_j \bar{\boldsymbol{\Delta}}_1(\varrho\beta_j, \varrho\beta'_j) \geq 0, \forall \boldsymbol{\beta}, \boldsymbol{\beta}'$. But Proposition 4 holds even when P is nonconcave on $\mathbb{R}_+$ and l is nonconvex.

The global convergence-rate results presented in this subsection are free of any regularity conditions on sparsity, sample size, initial point and design incoherence. High-dimensional learning algorithms may however show a better convergence rate when the problems under consideration are "regular" in a certain sense.

3.2. *Statistical accuracy.* To statisticians, the statistical accuracy of Bregman-surrogate algorithms with respect to a statistical truth (denoted by $\boldsymbol{\beta}^*$) is perhaps more meaningful than the optimization error to a certain local or global minimizer, since real world data are always noisy. Section 3.2.1 and Section 3.2.2 will study the statistical error of the final estimate $\hat{\boldsymbol{\beta}}$ and the t-th iterate $\boldsymbol{\beta}^{(t)}$, respectively, where combining the generalized Bregman calculus and the empirical process theory eases the treatment of a nonquadratic loss.

The techniques based on GBFs apply to a general problem (see, e.g., Theorem A.1 in Section A.18), but here we focus on the aforementioned sparse learning in the composite setting: $\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda)$, where $l(\boldsymbol{\beta}) = l_0(\boldsymbol{\eta}) = l_0(\boldsymbol{X}\boldsymbol{\beta})$ is directionally differentiable and $P_\Theta(\cdot; \lambda)$ is induced by a thresholding Θ via (48). Since l_0 is placed on $\boldsymbol{X}\boldsymbol{\beta}$, we include here a scaling parameter ϱ (often $\|\boldsymbol{X}\|_2$) in the penalty; this will yield a universal choice of

the regularization parameter λ that does not vary with the sample size. Throughout Section 3.2, we assume that ϱ satisfies $\varrho \geq \|\boldsymbol{X}\|_2$. Note that neither the loss nor the penalty needs to be convex or smooth.

Give any directionally differentiable ψ, the sequence of iterates is generated by

$$\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) := l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}). \tag{52}$$

Nonconvex iterative thresholding and LLA are particular instances.

First, we must characterize the notion of noise in this nonlikelihood setting, to take into account the randomness of samples. Assume l_0 is differentiable at point $\boldsymbol{X}\boldsymbol{\beta}^*$ (but not necessarily differentiable on all of $\mathbb{R}^n$) and define the *effective noise* by

$$\boldsymbol{\epsilon} = -\nabla l_0(\boldsymbol{X}\boldsymbol{\beta}^*). \tag{53}$$

(An alternative assumption is that $\delta l_0(\boldsymbol{X}\boldsymbol{\beta}^*; \boldsymbol{h})$ is a sub-Gaussian random variable with mean 0 and scale bounded by $c\sigma$ for any unit vector $\boldsymbol{h}$, but we will not pursue further in the current paper.)

Typically, $\mathbb{E}[\boldsymbol{\epsilon}]$ should be 0, and so $\nabla\{\mathbb{E}[l_0(\boldsymbol{X}\boldsymbol{\beta}^*)]\} = 0$ assuming the differentiation and expectation are exchangeable, which means the statistical truth makes the gradient of its risk vanish. For a GLM with y_i $(1 \leq i \leq n)$ following a distribution in the exponential family that has cumulant function b and canonical link function $g = (b')^{-1}$, the loss is then $l(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta}) = -\langle \boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta}\rangle + \langle \mathbf{1}, b(\boldsymbol{X}\boldsymbol{\beta})\rangle$ (cf. (9) with $\sigma = 1$), and so

$$\boldsymbol{\epsilon} = \boldsymbol{y} - g^{-1}(\boldsymbol{X}\boldsymbol{\beta}^*) = \boldsymbol{y} - \mathbb{E}(\boldsymbol{y}). \tag{54}$$

Our effective noise, as a joint outcome of the loss and the response, does not depend on the regularizer, and may differ from the raw noise. For example, under $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}^{\text{raw}}$, $l(\boldsymbol{\beta}) = l_{\text{Huber}}(\boldsymbol{r}) = \sum_{i:|r_i|\leq a\sigma} r_i^2/2 + \sum_{i:|r_i|>a\sigma}(a|r_i| - a^2\sigma^2/2)$ with $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}$ [27], simple calculation gives $\epsilon_i = \epsilon_i^{\text{raw}} 1_{|\epsilon_i^{\text{raw}}|\leq a\sigma} + a\sigma 1_{|\epsilon_i^{\text{raw}}|>a\sigma}$, which is bounded by $a\sigma$, thereby sub-Gaussian, no matter what distribution the raw noise follows. This nonparametricness is apparent for any l_0 that is (globally) Lipschitz, for example, the logistic deviance and hinge loss for classification.

In this section, we assume that $\boldsymbol{\epsilon}$ is a sub-Gaussian random vector with mean zero and scale bounded by σ, cf. Definition A.1, where ϵ_i are not required to be independent. Examples include Gaussian random variables and bounded random variables such as Bernoulli.

The support of $\boldsymbol{\beta}$ is denoted by $\mathcal{J}(\boldsymbol{\beta}) = \{j : \beta_j \neq 0\}$, and its cardinality is $J(\boldsymbol{\beta}) = |\mathcal{J}(\boldsymbol{\beta})| = \|\boldsymbol{\beta}\|_0$. We abbreviate $J(\boldsymbol{\beta}^*)$ to J^* and $J(\hat{\boldsymbol{\beta}})$ to $\hat{J}$. In sparse learning, $J^* \ll n \ll p$ is typically true. The sparsity suggests the possibility of obtaining a fast rate of convergence in statistical error. The following penalty induced by the hard-thresholding $\Theta_H(t;\lambda) = t1_{|t|>\lambda}$ by (48) turns out to play a key role in the analysis

$$P_H(t;\lambda) = (-t^2/2 + \lambda|t|)1_{|t|<\lambda} + (\lambda^2/2)1_{|t|\geq\lambda}. \tag{55}$$

An important fact is that $P_\Theta(t;\lambda) \geq P_H(t;\lambda)$ for any $t \in \mathbb{R}$ and any thresholding rule Θ. This is simply because in shrinkage estimation, any $\Theta(t;\lambda)$ with λ as the threshold is identical to zero as $t \in [0,\lambda)$ and is bounded above by the identity line for $t \geq \lambda$.

3.2.1. *Statistical accuracy of fixed-point solutions.* The finally obtained solutions from a Bregman surrogate algorithm can be described as the fixed points of g (recall (32)),

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}). \tag{56}$$

We denote the set by $\mathcal{F}$, and call such solutions the F*-estimators*. When the objective function is convex, an F-estimator is necessarily a globally optimal solution to the original problem by

Lemma 4, thus an M-estimator. In general, however, the lack of convexity and smoothness may make $\hat{\boldsymbol{\beta}}$ neither an M-estimator nor a Z-estimator [60], which poses new and intriguing challenges to statistical algorithmic analysis. It is also worth mentioning that another important class of "A-estimators" that have *alternative* optimality, typically arising from block coordinate descent (BCD) algorithms like in Example 3, can often be converted to F-estimators; see Section A.17.

Nicely, if the problem is regular, all F-estimators defined through g can achieve essentially the best statistical precision in possibly high dimensions. This is nontrivial since even f's locally optimal solutions do not all have the provable guarantee (cf. Remark 4). Theorem 3 and Theorem 4 below only make use of the weak idempotence property; another notable feature is that the conditions and conclusions below are *regardless* of the form of $\boldsymbol{\Delta}_\psi$.

THEOREM 3. *Suppose there exist $\delta > 0$, $\vartheta > 0$ and large enough $K \geq 0$ so that the following inequality holds for any $\boldsymbol{\beta} \in \mathbb{R}^p$:*

$$\begin{aligned}&\varrho^2\mathcal{L}_\Theta\mathbf{D}_2(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{X}\boldsymbol{\beta}^*) + \vartheta P_H(\varrho(\boldsymbol{\beta}-\boldsymbol{\beta}^*);\lambda) + P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda)\\ &\leq 2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{\beta},\boldsymbol{\beta}^*) + P_\Theta(\varrho\boldsymbol{\beta};\lambda) + K\lambda^2 J(\boldsymbol{\beta}^*),\end{aligned} \tag{57}$$

where $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta\wedge\vartheta)\vartheta}$ with A a sufficiently large constant. Then

$$\mathbf{D}_2(\boldsymbol{X}\hat{\boldsymbol{\beta}},\boldsymbol{X}\boldsymbol{\beta}^*) \leq \frac{2KA^2}{(\delta\wedge\vartheta)\delta\vartheta}\sigma^2 J^*\log(ep), \tag{58}$$

$$P_H(\varrho(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}^*);\lambda) \leq \frac{4KA^2}{(\delta\wedge\vartheta)\vartheta^2}\sigma^2 J^*\log(ep), \tag{59}$$

with probability at least $1 - Cp^{-cA^2}$, where C, c are positive constants.

Moreover, an *oracle inequality* [17, 31] can be built to justify the estimators even when $\boldsymbol{\beta}^*$ is not exactly sparse. Toward this goal, recall the notion of a pseudo-metric d (cf. Definition A.2), that is, d is nonnegative, symmetric, and satisfies the triangle inequality, and suppose without loss of generality that

$$\alpha d^2(\boldsymbol{\eta},\boldsymbol{\eta}') \leq \boldsymbol{\Delta}_{l_0}(\boldsymbol{\eta},\boldsymbol{\eta}') \leq Ld^2(\boldsymbol{\eta},\boldsymbol{\eta}'), \forall \boldsymbol{\eta},\boldsymbol{\eta}'$$

for some pseudo-metric d with $-\infty \leq \alpha \leq L \leq +\infty$. For regression $l(\boldsymbol{\beta}) = l_0(\boldsymbol{\eta}) = \|\boldsymbol{y}-\boldsymbol{\eta}\|_2^2/2$, $\alpha = L = 1 > 0$.

THEOREM 4. *Assume for given $\boldsymbol{\beta} \in \mathbb{R}^p$, there exist r: $0 \leq r < 1, \alpha r/L \geq 0$, positive δ, ϑ, and a large enough $K \geq 0$ so that*

$$\begin{aligned}&\varrho^2\mathcal{L}_\Theta\mathbf{D}_2(\boldsymbol{\beta},\boldsymbol{\gamma}) + \delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{X}\boldsymbol{\gamma}) + \vartheta P_H(\varrho(\boldsymbol{\beta}-\boldsymbol{\gamma});\lambda) + P_\Theta(\varrho\boldsymbol{\beta};\lambda)\\ &\leq (1+\frac{\alpha}{L}r)\boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\gamma}) + P_\Theta(\varrho\boldsymbol{\gamma};\lambda) + K\lambda^2 J(\boldsymbol{\beta})\end{aligned} \tag{60}$$

for any $\boldsymbol{\gamma} \in \mathbb{R}^p$, where $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta\wedge\vartheta)\vartheta}$ with A a sufficiently large constant. The oracle inequality below holds for some constant $C > 0$,

$$\mathbb{E}\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*) \leq \mathbb{E}\Big\{\Big(\frac{1+r}{1-r}\Big)^2\boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \frac{(1+r)KA^2}{(1-r)(2\vartheta\wedge\delta)\vartheta}\sigma^2 J(\boldsymbol{\beta})\log(ep)\Big\} + \frac{C(1+r)}{(1-r)(2\vartheta\wedge\delta)}\sigma^2. \tag{61}$$

Compared with (57) which fixes $\boldsymbol{\gamma}$ at $\boldsymbol{\beta}^*$, (60) has $(1+\frac{\alpha}{L}r)\boldsymbol{\Delta}_l$ in place of $2\bar{\boldsymbol{\Delta}}_l$ as the first term on the right-hand side. Nonrigorously, these conditions ask $2\bar{\boldsymbol{\Delta}}_l$ or $(1+\frac{\alpha}{L}r)\boldsymbol{\Delta}_l$ to dominate $\varrho^2\mathcal{L}_\Theta\mathbf{D}_2$ in a restricted sense; Remark 2 argues that (60) is not technically demanding compared with many other regularity conditions in the literature.

When $r = 0$, the multiplicative constant proceeding $\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ in (61) is as small as 1, resulting in a sharp oracle inequality [31]. If one sets $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in (61), the Bregman error $\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$ is of the order $\sigma^2 J^* \log(ep)$ for any thresholding (when δ, ϑ, K are treated as constants). But the bias term $\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ or $\boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}^*)$ helps to handle *approximately* sparse signals: when $\boldsymbol{\beta}^*$ contains a number of small nonzero elements, rather than taking $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, a reference $\boldsymbol{\beta}$ with a reduced support will yield an even smaller error bound benefiting from the bias-variance tradeoff.

Unlike the optimization error bounds, the statistical error bounds never vanish (unless $\sigma \to 0$). We can similarly analyze the set of global minimizers, in which case the term $\varrho^2 \mathcal{L}_\Theta \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^*)$ is dropped from the regularity conditions, but the error bounds remain of the same order (cf. Remark A.1 in Section A.12). In fact, for sparse GLMs, by Theorem 1, the rate $\sigma^2 J^* \log(ep)$ is essentially minimax optimal (thus unbeatable) up to a logarithmic factor.

REMARK 2 (Regularity condition comparison). *The GBF-based regularity conditions (57), (60) are no more demanding than some commonly used regularity conditions. Assume that P_Θ is subadditive: $P_\Theta(t+s) \le P_\Theta(t) + P_\Theta(s)$, which holds when it is concave on $\mathbb{R}_+$. Let $\mathcal{J} = \mathcal{J}(\boldsymbol{\beta})$, $J = |\mathcal{J}(\boldsymbol{\beta})|$, $\boldsymbol{\gamma} = \boldsymbol{\beta}' - \boldsymbol{\beta}$. Then, from $P_\Theta(\varrho\boldsymbol{\beta}'_{\mathcal{J}}; \lambda) - P_\Theta(\varrho\boldsymbol{\beta}_{\mathcal{J}}; \lambda) \le P_\Theta(\varrho(\boldsymbol{\beta}' - \boldsymbol{\beta})_{\mathcal{J}}; \lambda)$ and $P_\Theta(\varrho\boldsymbol{\beta}'_{\mathcal{J}^c}; \lambda) = P_\Theta(\varrho(\boldsymbol{\beta}' - \boldsymbol{\beta})_{\mathcal{J}^c}; \lambda)$, (60) is implied by $P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \vartheta P_H(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \mathcal{L}_\Theta \mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \delta\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2/2 \le (2-\varepsilon)\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + K\lambda^2 J + P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda) - \vartheta P_H(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda)$, or $(1+\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \mathcal{L}_\Theta \mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \delta\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2/2 \le (2-\varepsilon)\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + K\lambda^2 J + (1-\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda)$ since $P_H \le P_\Theta$.*

To get more intuition, let $l(\boldsymbol{\beta}) = \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2/2$. Then the above condition simplifies to $(1+\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \mathcal{L}_\Theta\|\varrho\boldsymbol{\gamma}\|_2^2/2 \le (2-\varepsilon')\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2/2 + K\lambda^2 J + (1-\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda)$ with $\varepsilon' = \varepsilon + \delta$, or the following sufficient condition (with K redefined) for all $\boldsymbol{\gamma} \in \mathbb{R}^p$:

$$(1+\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \frac{\mathcal{L}_\Theta}{2}\|\varrho\boldsymbol{\gamma}\|_2^2 \le K\sqrt{J}\lambda\|\boldsymbol{X}\boldsymbol{\gamma}\|_2 + (1-\vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda). \tag{62}$$

For lasso, where $P_\Theta(\boldsymbol{\beta}; \lambda) = \lambda\|\boldsymbol{\beta}\|_1$, there is a rich collection of regularity conditions in the literature. In this convex case, $\mathcal{L}_\Theta = 0$ and ϱ can be arbitrarily large. (62) reduces to (with ϑ and K redefined and λ canceled)

$$(1+\vartheta)\varrho\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \le K\sqrt{J}\|\boldsymbol{X}\boldsymbol{\gamma}\|_2 + \varrho\|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1, \forall\boldsymbol{\gamma} \tag{63}$$

for some $K \ge 0, \vartheta > 0$. Taking $\varrho = c\|\boldsymbol{X}\|_2$ results in scale invariance with respect to $\boldsymbol{X}$. Let's compare (63) with the restricted eigenvalue (RE) condition and the compatibility condition [7, 59]. For given $\mathcal{J}$, the two conditions assume that there exist positive numbers κ, ϑ_{RE} such that $J\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \ge \kappa\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1^2$ (compatibility) or more restrictively, $\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \ge \kappa\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2$ (RE), for all $\boldsymbol{\gamma} : (1+\vartheta_{RE})\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \ge \|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1$. Therefore, $(1+\vartheta)\varrho\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \le K\sqrt{J}\|\boldsymbol{X}\boldsymbol{\gamma}\|_2 \vee \varrho\|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1$ with $K = (1+\vartheta_{RE})/(\varrho\sqrt{\kappa})$, $\vartheta = \vartheta_{RE}$. That is, the RE-type conditions are more demanding than (63) (and (60)). Another popular set of regularity conditions is based on restricted strong convexity (RSC). Under a version of RSC condition (and assuming f is differentiable), [36, Theorem 1] showed that $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2$ has a bound of order $\sigma^2(J^* \log p)/n$ for any stationary point $\tilde{\boldsymbol{\beta}}$. In the lasso case, the condition becomes $\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \ge \alpha\|\boldsymbol{\gamma}\|_2^2 - \tau\log p\|\boldsymbol{\gamma}\|_1^2/n$ for some constant $\alpha > 0$ and $\tau \ge 0$, from which it follows that for any $\boldsymbol{\gamma} : (1+\vartheta_{RE})\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \ge \|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1$, $\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \ge \alpha\|\boldsymbol{\gamma}\|_2^2 - \tau(2+\vartheta_{RE})^2\frac{\log p}{n}\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1^2 \ge \alpha\|\boldsymbol{\gamma}\|_2^2 - \tau(2+\vartheta_{RE})^2\frac{J\log p}{n}\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2 \ge \kappa'\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2$, where $\kappa' = \alpha - \tau(2+\vartheta_{RE})^2(J\log p/n)$. Therefore, when $n \gg J\log p$, RSC implies RE and so is more restrictive than (63). See Remark A.1 in Section A.12 for an extension to general penalties.

REMARK 3 (Technical treatment). *A big difference between our work and [36] is that the latter enforces an ℓ_1-type side constraint, for example, $\|\boldsymbol{\beta}\|_1 \le R$, in addition to the sparsity-inducing penalty P. The use of the constraint is a necessary ingredient of the proofs and the constraint parameter R appears in the minimum sample size condition and the error bounds implicitly. However, few practically used algorithms seem to include such an additional ℓ_1 constraint.*

Our analysis does not need any side constraint, and the resulting error bounds and the oracle inequality hold with no minimum sample size requirement. In fact, in dealing with a general penalty that may be nonconvex, our treatment of the stochastic term is distinctive from the conventional "ℓ_1 fashion" via Hölder's inequality: $\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle \le \|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty \|\boldsymbol{\beta}\|_1$ (see, e.g., [10, 7, 37]). More concretely, applying the union bound to $\|\boldsymbol{X}^\top \boldsymbol{\epsilon}\|_\infty$ will lead to a further upper bound $\|\boldsymbol{\beta}\|_2^2 + P(\boldsymbol{\beta};\lambda)$ up to multiplicative factors [36], while we can bound $\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle$ by the sum of $\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2/a$ and a light penalty $P_H(\boldsymbol{\beta};\lambda)/b$ for any $a, b > 0$, with a proper choice of λ.

REMARK 4 (Fixed points vs. local minimizers). *Targeting at the fixed points of the Bregman surrogate instead of the local minimizers of the original objective seems more reasonable from a statistical perspective. Certainly, if f is smooth, $\mathcal{F}$ contains more valid solutions (cf. Lemma 4). But a more important reason is that $\mathcal{F}$ can adaptively exclude bad local solutions for some statistical learning problems with severe nonsmoothness and nonconvexity.*

For instance, each bridge ℓ_q-penalty ($q : 0 \le q < 1$) [22] determines a thresholding Θ_q, which is however the solution for infinitely many penalties; picking the particular one constructed from (48) that is the lowest and directionally differentiable [49], one can repeat the analysis in Theorems 3, 4 to show provable guarantees for all the fixed points of the iterative Θ_q procedure. In contrast, as pointed out by [36], the original optimization problem may contain "faulty" local minimizers. In fact, when $q = 0$, the ℓ_0-penalized problem $\min_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2/2 + (\lambda^2/2)\|\boldsymbol{\beta}\|_0$ (not directionally differentiable) always *has $\mathbf{0}$ as a local minimizer which is however a poor estimator as $\boldsymbol{\beta}^*$ is large. Switching to the surrogate's fixed points successfully addresses the issue: $\hat{\boldsymbol{\beta}} = \mathbf{0}$ is a valid fixed point only when $\boldsymbol{X}^\top \boldsymbol{y}$ is properly small: $\|\boldsymbol{X}^\top \boldsymbol{y}\|_\infty \le \lambda$, or the true signal is inconsequential relative to the maximum noise level.*

3.2.2. *Statistical analysis of the iterates from Bregman surrogates.* We show a nice result for (52) in the composite setting: under a regularity condition similar to those in Section 3.2.1, with high probability, the t-th iterate can approach the statistical target within the desired precision geometrically fast, even when $p > n$. Specifically, we add a mild multiple of $\boldsymbol{\Delta}_\psi$ to the left-hand side of (57) and assume that for some $\delta > 0$, $\varepsilon > 0$, $\vartheta > 0$ and large $K \ge 0$,

$$\begin{aligned} &\varepsilon \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}) + \delta \mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}^*) + \vartheta P_H(\varrho(\boldsymbol{\beta} - \boldsymbol{\beta}^*);\lambda) + P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda) \\ &\le (2\bar{\boldsymbol{\Delta}}_l - \varrho^2 \mathcal{L}_\Theta \mathbf{D}_2)(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + P_\Theta(\varrho\boldsymbol{\beta};\lambda) + K\lambda^2 J(\boldsymbol{\beta}^*), \forall \boldsymbol{\beta} \end{aligned} \tag{64}$$

and ψ is differentiable for simplicity. Recall that (39) in Proposition 2 requires $2\bar{\boldsymbol{\Delta}}_f$ to dominate $\varepsilon\boldsymbol{\Delta}_\psi$; (64) gives a large-$p$ extension of it.

THEOREM 5. *Under the above regularity condition, for $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta \wedge \vartheta)\vartheta}$ with A sufficiently large and $\kappa = 1/(1+\varepsilon)$, we have*

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}) \le \kappa^t \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(0)}) + \frac{\kappa}{1-\kappa}(K\lambda^2 J^* - \min_{1 \le s \le t} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(s)}, \boldsymbol{\beta}^{(s-1)})) \tag{65}$$

for any $t \ge 1$ with probability at least $1 - Cp^{-cA^2}$, where C, c are universal positive constants.

The error measure $\mathbf{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)})$ in (65) has $\boldsymbol{\beta}^*$ as its first argument and differs from the $\mathbf{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*)$ used in (61). According to the proof, (64) only needs to hold for $\boldsymbol{\beta} = \boldsymbol{\beta}^{(s)}$ $(0 \le s \le t)$, and so different starting values may give different values of κ. With $\mathbf{\Delta}_\psi \ge 0$ (which can be realized by stepsize control), the fast converging statistical error to $\mathcal{O}(\sigma^2 J^* \log(ep))$ implies that over-optimization may be unnecessary. As an example, consider the iterative thresholding procedures with $\mathbf{\Delta}_l \le L\mathbf{D}_2$ and $\varrho^2 > L$. Then (65) yields

$$\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(t)}\|_2^2 \le \kappa^t \frac{\varrho^2}{\varrho^2 - L}\|\boldsymbol{\beta}^* - \boldsymbol{\beta}^{(0)}\|_2^2 + \frac{2\kappa K}{(1-\kappa)(\varrho^2 - L)}\lambda^2 J^*.$$

So it is possible to terminate the iterative algorithm before full computational convergence without sacrificing much statistical accuracy. The simulations in Section C.2 support this point.

REMARK 5. *Theorem 5 reveals the fast decay of the* direct *statistical error between $\boldsymbol{\beta}^{(t)}$ and $\boldsymbol{\beta}^*$. [1] and [36] argued a similar point for gradient descent type algorithms, in a somehow indirect manner: (i) $\boldsymbol{\beta}^{(t)}$ can approach any globally optimal solution $\tilde{\boldsymbol{\beta}}$ geometrically fast in computation under a combination of an RSC condition and an RSM condition, and (ii) under some regularity conditions, every local minimum point is close enough to the authentic $\boldsymbol{\beta}^*$. In the RSC condition for (i), the factor proceeding the dominant term $\bar{\mathbf{\Delta}}_l$ is* 1 *(there are two different sets of RSC conditions used in Theorem 1 and Theorem 3 of [36], the factor α_1 in the second set corresponding to* half *of the α_1 used in the first set). But* (64) *allows it to be* 2*. Moreover, Theorem 5 does not need the extra RSM condition and applies to a broader class of algorithms. For example, we can show that the statistical error of the LLA algorithm reduces at a linear rate to the desired precision under some regularity conditions; see Proposition 5 and Lemma A.7 in Section A.16.*

4. Two acceleration schemes for generalized Bregman surrogates. How to accelerate first-order algorithms without incurring much additional cost per iteration has lately attracted lots of attention in big data applications. In convex optimization, Nesterov's momentum techniques prove to be quite effective in that the rate of convergence can be improved from $\mathcal{O}(1/t)$ to $\mathcal{O}(1/t^2)$, which is optimal when using first-order methods on smooth problems [41, 57, 4, 32]. This section attempts to extend Nesterov's *first* and *second* accelerations [39, 40] to Bregman-surrogate algorithms. With a possible lack of smoothness or convexity, carefully choosing the relaxation parameters and step sizes is the key, and we will see the benefit of maximizing a quantity $R_t/(\theta_t^2 \rho_t)$ at the t-th iteration, with R_t appropriately defined via generalized Bregman notation. We consider the following two broad scenarios to devise the acceleration schemes.

Scenario 1. $g(\boldsymbol{\beta}; \boldsymbol{\gamma}) = f(\boldsymbol{\beta}) - \mathbf{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \rho\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma})$. This surrogate family includes gradient descent type algorithms. Often, if $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \mathbf{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is easy to solve, so is $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) + \mathbf{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \rho\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma})$, in which case $\psi_0 = -\psi$.

Scenario 2. $g(\boldsymbol{\beta}; \boldsymbol{\gamma}) = f(\boldsymbol{\beta}) - \mathbf{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}) + \rho\mathbf{\Delta}_\phi(\boldsymbol{\beta}, \boldsymbol{\gamma})$. This gives a more general class than the first one.

This section assumes that f, ψ_0, ϕ, $\mathbf{\Delta}_{\psi_0}(\cdot, \boldsymbol{\gamma})$, $\mathbf{\Delta}_{\psi_0}(\cdot, \boldsymbol{\gamma})$ are directionally differentiable given any $\boldsymbol{\gamma}$. We introduce a convenient notation $\mathbf{C}_\psi$ defined for any ψ as follows

$$\mathbf{C}_\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \theta) = \theta\psi(\boldsymbol{\alpha}) + (1-\theta)\psi(\boldsymbol{\beta}) - \psi(\theta\boldsymbol{\alpha} + (1-\theta)\boldsymbol{\beta}), \tag{66}$$

where $0 \le \theta \le 1$. Like $\mathbf{\Delta}$, $\mathbf{C}$ is a linear operator of ψ and its nonnegativity means convexity. Some connections between $\mathbf{\Delta}$ and $\mathbf{C}$ are given below.

LEMMA 5. *Let ψ be directionally differentiable. (i) $\mathbf{C}_\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\theta) = (1-\theta)\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\alpha}) - \boldsymbol{\Delta}_\psi(\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta},\boldsymbol{\alpha})$ for any $\boldsymbol{\alpha},\boldsymbol{\beta}$ and $\theta\in[0,1]$. (ii) $\mathbf{C}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})} = \mathbf{C}_\psi$ if ψ is differentiable at $\boldsymbol{\alpha}$.*

An acceleration scheme of the second kind. Scenario 2 is of our primary interest since it applies more broadly. Below, we modify the surrogate and define an iterative algorithm (not a descent method) that involves three sequences $\boldsymbol{\alpha}^{(t)}$, $\boldsymbol{\beta}^{(t)}$, $\boldsymbol{\gamma}^{(t)}$ starting at $\boldsymbol{\alpha}^{(0)} = \boldsymbol{\beta}^{(0)}$:

$$\boldsymbol{\gamma}^{(t)} = (1-\theta_t)\boldsymbol{\beta}^{(t)} + \theta_t\boldsymbol{\alpha}^{(t)} \tag{67a}$$

$$\boldsymbol{\alpha}^{(t+1)} = \operatorname{argmin} f(\boldsymbol{\beta}) - \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)}) + \mu_0\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)}) + \theta_t\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)}) \tag{67b}$$

$$\boldsymbol{\beta}^{(t+1)} = (1-\theta_t)\boldsymbol{\beta}^{(t)} + \theta_t\boldsymbol{\alpha}^{(t+1)}, \tag{67c}$$

for some $\mu_0 \geq 0$, $\theta_t \in (0,1]$, $\rho_t > 0$ ($\forall t \geq 0$), to be chosen later. Notice the extra GBF term $\mu_0\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\gamma}^{(t)})$ in (67b) in addition to $\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})$. The design of relaxation parameters θ_t and inverse step size parameters ρ_t, μ_0 holds the key to acceleration. Let

$$\bar{\psi}_0 = \psi_0 - \mu_0\phi. \tag{68}$$

We advocate the following line search criterion

$$\begin{aligned} R_t :=\ & \theta_t^2\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\alpha}^{(t)}) - \boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)}) \\ & + \mathbf{C}_{f(\cdot)-\boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot,\boldsymbol{\gamma}^{(t)})}(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\beta}^{(t)},\theta_t) \geq 0, \end{aligned} \tag{69a}$$

$$\frac{\theta_t^2}{1-\theta_t} = \frac{\theta_{t-1}(\rho_{t-1}\theta_{t-1}+\mu_0)}{\rho_t},\ t \geq 1. \tag{69b}$$

The update of the relaxation parameter involves ρ and μ as well.

Theorem 6 presents two error bounds without assuming convexity or smoothness, and shows in general the reasonability of (69a).

THEOREM 6. *Let ρ_t be any positive sequence. Consider the algorithm defined by (67a)–(67c) and (69b). Let $\mathcal{E}_t(\boldsymbol{\beta}) = \boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)}) + \boldsymbol{\Delta}_{f(\cdot)-\boldsymbol{\Delta}_{\psi_0}(\cdot,\boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)}) + (\mu_0\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\gamma}^{(t)})-\phi(\cdot)} + \theta_t\rho_t\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})-\phi(\cdot)})(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})$.*

(i) When $\mu_0 = 0$, for any $\boldsymbol{\beta}$ and $T \geq 0$,

$$\begin{aligned} & \frac{f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta})}{\theta_T^2\rho_T} + T\cdot\operatorname*{avg}_{0\leq t\leq T}\frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t\rho_t} + T\cdot\operatorname*{avg}_{0\leq t\leq T}\frac{R_t}{\theta_t^2\rho_t} \\ & \leq \boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(0)}) - \boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(T+1)}) + \frac{1-\theta_0}{\theta_0^2\rho_0}\big[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})\big]. \end{aligned} \tag{70}$$

(ii) Moreover, given any $\mu_0 \geq 0$,

$$\begin{aligned} & f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \theta_T^2(\rho_T + \frac{\mu_0}{\theta_T})\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(T+1)}) \\ & + \Sigma_{t=0}^T\big(\Pi_{s=t+1}^T(1-\theta_s)\big)(R_t + \theta_t\mathcal{E}_t(\boldsymbol{\beta})) \\ & \leq \Big(\prod_{t=1}^T(1-\theta_t)\Big)\big[(1-\theta_0)(f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})) + \theta_0^2\rho_0\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\beta}^{(0)})\big] \end{aligned} \tag{71}$$

for all $\boldsymbol{\beta}$ and $T \geq 0$, where by convention, $\prod_{s=l}^u a_s = 1$ as $l > u$.

First, we make a discussion of the results for convex optimization. Assume $\boldsymbol{\Delta}_\phi \geq \sigma \mathbf{D}_2$ for some $\sigma > 0$. With the additional knowledge that $f(\cdot) - \boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot, \boldsymbol{\gamma}^{(t)})$ is convex and $\boldsymbol{\Delta}_{\bar{\psi}_0} \leq L_{\bar{\psi}_0} \mathbf{D}_2$ for some $L_{\bar{\psi}_0} \geq 0$, (69a) is implied by

$$\theta_t^2(\rho_t - L_{\bar{\psi}_0}/\sigma)\boldsymbol{\Delta}_\phi(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) + (1-\theta_t)\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \geq 0. \tag{72}$$

So when f is convex, criterion (69) is satisfied by $\rho_t = \rho \geq L_{\bar{\psi}_0}/\sigma, \psi_0 = f, \mu_0 = 0$ and $\theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$, degenerating to Nesterov's second method [40, 57], and the convergence rate is of order $\mathcal{O}(1/T^2)$ according to (70) and (75). The second conclusion tells more when strong convexity (or restricted strong convexity) arises. Given a convex f satisfying $\mu \mathbf{D}_\phi \leq \boldsymbol{\Delta}_f \leq L \mathbf{D}_\phi$ with $0 < \mu \leq L$ and ϕ differentiable, taking $\psi_0 = f, \mu_0 = \mu$, and $\rho_t = L - \mu$ ensures $\mathcal{E}_t(\boldsymbol{\beta}) = \boldsymbol{\Delta}_{f-\mu\phi}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \boldsymbol{\Delta}_{f(\cdot)-\boldsymbol{\Delta}_f(\cdot,\boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(t+1)}) \geq \boldsymbol{\Delta}_{f-\mu\phi}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) \geq 0$ and $R_t \geq \theta_t^2 \rho_t \mathbf{D}_\phi(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) - \boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) \geq \theta_t^2(\rho_t + \mu_0 - L)\sigma \mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) = 0$. According to (69b), the following choice

$$\theta_t = \theta_0 = \frac{2}{\sqrt{4\kappa - 3} + 1} \text{ with } \kappa = L/\mu \tag{73}$$

suffices, and the optimization problem to solve in (67b) becomes

$$\min f(\boldsymbol{\gamma}^{(t)}) + \delta f(\boldsymbol{\beta}; \boldsymbol{\beta} - \boldsymbol{\gamma}^{(t)}) + \mu \mathbf{D}_\phi(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \frac{2(L-\mu)}{\sqrt{4\kappa - 3} + 1} \mathbf{D}_\phi(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(t)}). \tag{74}$$

From (71), both $f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta})$ and $\mathbf{D}_\phi(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(T+1)})$ enjoy a linear convergence with rate parameter $\frac{\sqrt{4\kappa-3}-1}{\sqrt{4\kappa-3}+1}$, or an iteration complexity of $\mathcal{O}(\sqrt{\kappa}\log(1/\epsilon))$, significantly faster than $\mathcal{O}(\kappa \log(1/\epsilon))$ in Proposition 2. Hence (67), (69) can achieve rate-optimality in various convex scenarios. To the best of our knowledge, this is the first "all-in-one" form of the *second* acceleration that adapts.

The proposed algorithm can even go beyond convexity. As a demonstration, let us apply the acceleration to the iterative quantile-thresholding procedure (cf. Example 2) for solving the feature screening problem: $\min l(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$ s.t. $\|\boldsymbol{\beta}\|_0 \leq q$, which is nonconvex. Here, q is bounded above by p but may be larger than n. Take $\phi = \|\cdot\|_2^2/2$, $\mu_0 = 0$ and $\psi_0(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \mathcal{L}\phi(\boldsymbol{\beta})$ for some $\mathcal{L} \geq 0$. Given any $s \leq p$ and $\boldsymbol{X}$, define the restricted isometry number $\rho_+(s)$ [12] that satisfies $\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \leq \rho_+(s)\|\boldsymbol{\beta}\|_2^2, \forall \boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \leq s$, which can be much smaller than $\|\boldsymbol{X}\|_2^2$ as s is small.

COROLLARY 2. *Assume q is set larger than the target $\|\boldsymbol{\beta}^*\|_0$ with the ratio denoted by r. Then for any $\mathcal{L} \geq \rho_+(2q)/\sqrt{r}$, there exists a* universal *ρ_t ($\rho_t = \rho_+(2q)(1 - 1/\sqrt{r})$, say), thereby $\theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$, such that the accelerated iterative quantile-thresholding according to (67a)–(67c) satisfies $l(\boldsymbol{\beta}^{(T+1)}) - l(\boldsymbol{\beta}^*) + \min_{0 \leq t \leq T} \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)}) \leq A/T^2$ for all $T \geq 0$, where A is independent of T.*

The proof of the corollary shows the power of an *accumulative* R_t-control, and applies more generally: if the objective function $f(\boldsymbol{\beta})$, possibly nonconvex, can be written as the sum of a convex function $l(\boldsymbol{\beta})$ with $\boldsymbol{\Delta}_l \leq L\mathbf{D}_2$ and a function $P(\boldsymbol{\beta})$ that can be lifted: $\boldsymbol{\Delta}_P + \mathcal{L}_0 \mathbf{D}_2 \geq 0$ for some finite $\mathcal{L}_0 \geq 0$, then one can utilize a ψ_0 as $l - 0.6\mathcal{L}_0\|\cdot\|_2^2$ and a universal ρ_t to fulfill $T \cdot \text{avg}_{t \leq T} R_t/(\theta_t^2 \rho_t) \geq 0$ in (70) (although not every R_t is necessarily nonnegative) so as to attain an $\mathcal{O}(1/T^2)$ error bound. See Remark A.3 in Section A.14.

Of course, a time-varying ρ_t can provide finer control, and the theorem does not limit ρ_t to be constant. In fact, under $\mu_0 = 0$, as long as $\rho_t/\rho_{t-1} \geq 1 - (at + ab + 1)/(t + b - 1)^2$ $(t \geq 1)$ for some constants a, b: $a > -2, b \geq a + 1$, induction based on (69b) gives $\theta_t \leq (a+2)/(t+b)$

and $\sum_{t=0}^{T}\rho_T/(\rho_t\theta_t) \geq (T+c_1)^2/(a+2)^2 + c_2$ (with constants c_i dependent on a, b) for any $t \geq 1$, from which it follows that

$$\theta_T^2 = \mathcal{O}(1/T^2) \text{ and } T \cdot \operatorname{avg}_{0\leq t\leq T}(1/(\rho_t\theta_t)) \geq \mathcal{O}(T^2/\rho_T). \tag{75}$$

Now, under $R_t \geq 0$ or just $\sum_{t=0}^{T} R_t/(\theta_t^2\rho_t) \geq 0$, (70) gives $f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \min_{0\leq t\leq T}\mathcal{E}_t(\boldsymbol{\beta}) \leq \mathcal{O}(\rho_T/T^2)$ for any $\boldsymbol{\beta}$. Typically, (69a) involves a line search. If the condition fails for the current value of ρ_t, one can set $\rho_t = \alpha\rho_t$ for some $\alpha > 1$ and recalculate θ_t, $\boldsymbol{\gamma}^{(t)}$, $\boldsymbol{\alpha}^{(t+1)}$ and $\boldsymbol{\beta}^{(t+1)}$ according to (69b) and (67) to verify it again. In implementation, it is wise to limit the number of searches at each iteration (denoted by M) to control the per-iteration complexity. If (69a) does not hold after m times of search, we simply pick the ρ_t that gives the largest $R_t/(\theta_t^2\rho_t)$ based on Theorem 6. Some details are in Algorithm B.1. In simulation studies, letting $M = 3$, $\alpha = 2$ already shows excellent performance; see Figure C.5 and Figure C.6.

An acceleration scheme of the first kind. For the algorithms falling into Scenario 1, we can alternatively consider two sequences of iterates generated by

$$\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + \{\rho_{t-1}\theta_t(1-\theta_{t-1})/(\rho_{t-1}\theta_{t-1}+\mu_0)\}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}), \tag{76a}$$

$$\boldsymbol{\beta}^{(t+1)} = \arg\min f(\boldsymbol{\beta}) - \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \mu_0\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \rho_t\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}), \tag{76b}$$

for some $\mu_0 \geq 0$, $\theta_t \in (0,1]$, $\rho_t > 0$ for all $t \geq 0$, and we force $\boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}^{(0)}$. (76a), (76b) give a new first type acceleration, and notably, the novel update of $\boldsymbol{\gamma}^{(t)}$ involves ρ_{t-1}. When $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\gamma}^{(t)}$ one stops the algorithm and obtains a fixed point with provable statistical guarantees as shown in Section 3.2.1.

Similar to (68), let $\bar{\psi}_0 = \psi_0 - \mu_0\|\cdot\|_2^2/2$. Define the line search criterion

$$R_t := (\rho_t\mathbf{D}_2 - \boldsymbol{\Delta}_{\bar{\psi}_0})(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \geq 0, \tag{77a}$$

$$\frac{\theta_t^2}{1-\theta_t} = \frac{\theta_{t-1}(\rho_{t-1}\theta_{t-1}+\mu_0)}{\rho_t},\ \theta_t \geq 0,\ \rho_t > 0,\ t \geq 1. \tag{77b}$$

Note that R_t is defined differently from (69a). The following theorem reveals the importance of maximizing R_t in each iteration step when performing possibly nonconvex optimization.

THEOREM 7. *Given any $\rho_t > 0$ ($t \geq 0$), consider the algorithm defined by (76a), (76b), and (77b). Let $\mathcal{E}_t(\boldsymbol{\beta}) = \boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \{\mathbf{C}_{f(\cdot)-\boldsymbol{\Delta}_{\psi_0}(\cdot,\boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t) + \boldsymbol{\Delta}_{f(\cdot)-\boldsymbol{\Delta}_{\psi_0}(\cdot,\boldsymbol{\gamma}^{(t)})}(\theta_t\boldsymbol{\beta} + (1-\theta_t)\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)})\}/\theta_t$.*

(i) When $\mu_0 = 0$, we have

$$\frac{f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta})}{\theta_T^2\rho_T} + T\cdot\operatorname*{avg}_{0\leq t\leq T}\frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t\rho_t} + T\cdot\operatorname*{avg}_{0\leq t\leq T}\frac{R_t}{\theta_t^2\rho_t}$$

$$\leq \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) + \frac{1-\theta_0}{\theta_0^2\rho_0}\left[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})\right] \text{ for any } \boldsymbol{\beta} \text{ and } T \geq 0.$$

(ii) Moreover, given any $\mu_0 \geq 0$, for all $\boldsymbol{\beta}$ and $T \geq 0$,

$$f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \theta_T^2(\rho_T + \frac{\mu_0}{\theta_T})\mathbf{D}_2(\boldsymbol{\beta}, (\boldsymbol{\gamma}^{(T+1)} - (1-\theta_{T+1})\boldsymbol{\beta}^{(T+1)})/\theta_{T+1})$$

$$+ \Sigma_{t=0}^{T}\big(\Pi_{s=t+1}^{T}(1-\theta_s)\big)(R_t + \theta_t\mathcal{E}_t(\boldsymbol{\beta}))$$

$$\leq \Big(\prod_{t=1}^{T}(1-\theta_t)\Big)\left[(1-\theta_0)(f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})) + \theta_0^2\rho_0\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)})\right].$$

Again, the new proposal of the iterate and parameter updates adapts to various situations, with μ_0 (which can be a sequence μ_t, cf. Remark A.2) measuring the degree of convexity (or restricted convexity in a nonconvex composite problem). For example, when f is convex and L-strongly smooth, $\mu_0 = 0$, $\rho_t = L$, $\psi_0 = f$, and $\theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$ make (77) hold, corresponding to Nesterov's first method. Interestingly, if f is μ-strongly convex, the associated standard momentum update $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + \theta_t(\theta_{t-1}^{-1} - 1)(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)})$ only attains a linear rate at $1 - 1/\kappa$ ($\kappa = L/\mu$) (cf. Remark A.4), showing *no* theoretical advantage over the plain gradient descent. (76) fixes the issue: with $\mu_0 = \mu$, $\rho_t = L - \mu$, $\theta_t = 2/(\sqrt{4\kappa - 3} + 1)$, an accelerated linear rate parameter is obtained as $(\sqrt{4\kappa - 3} - 1)/(\sqrt{4\kappa - 3} + 1)(\leq 1 - \sqrt{3/(4\kappa)})$. (When μ_0 is unknown, (76b) based on the split $L = \rho_t + \mu_t$ is still advantageous over the classical acceleration with $\rho_t = L$.) We proved these error bounds by use of GBFs, which is perhaps more straightforward than Nesterov's ingenious proof based on the notion of estimate sequence, and more importantly, (76), (77) provide a universal "all-in-one" form, instead of separate schemes in different situations [41].

Theorem 7 accommodates diverse choices of the parameters $\psi_0, \mu_0, \rho_t, \theta_t$ and is motivating in the nonconvex composite setup. Consider, for example, $\min f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2 + P_\Theta(\varrho\boldsymbol{\beta}; \lambda)$. Because the objective is nonconvex when $p > n$ and $\mathcal{L}_\Theta > 0$, how to accelerate the associated iterative thresholding procedure is an unconventional problem. From the studies in Section 3.2, we have learned that a sparsity-inducing penalty with a properly large threshold to suppress the noise can result in strong convexity in a restricted sense. We can then use a surrogate $f(\boldsymbol{\beta}) + (\rho\mathbf{D}_2 - \boldsymbol{\Delta}_{\psi_0})(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$ where $\psi_0(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2 - \varrho^2\mathcal{L}_\Theta\|\boldsymbol{\beta}\|_2^2/2$ and $\mu_0 = 0$. Since $f(\cdot) - \boldsymbol{\Delta}_{\psi_0}(\cdot, \boldsymbol{\gamma})$ is convex (cf. Lemma A.3), $\mathcal{E}_t(\boldsymbol{\beta}) \geq \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)})$. Moreover, thanks to the sparsity in $\boldsymbol{\beta}^{(t)}$, and thus $\boldsymbol{\gamma}^{(t)}$, $\boldsymbol{X}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\gamma}^{(t)})$ involves just a small number of features. So with an incoherent design, a properly small ϱ can make $\boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \geq 0$. Now, taking a constant ρ_t as large as, for instance, $\|\boldsymbol{X}\|_2^2 - \varrho^2\mathcal{L}_\Theta$, may yield a convergence rate of order $\mathcal{O}(1/t^2)$. (Actually, linear convergence may result from the restricted strong convexity under some regularity conditions.) More generally, different ρ_t's are allowed in the theorem: (75) is still secured with just, say, $\rho_t/\rho_{t-1} \geq 1 - (t+3)/(t+1)^2$. A line search can be used to determine a proper sequence ρ_t; see Algorithm B.2 for more details.

The proposed accelerations of the first kind and of the second kind can be utilized in a wide range of problems. Because they are momentum based, the original algorithms need not be substantially modified to have an improved iteration complexity, and the two theorems proved in this section apply in any dimensions with no design coherence restrictions. Another delightful fact is that our "all-in-one" forms update the iterates adaptively according to the degree of convexity $\mu_0 \geq 0$, which can be relaxed to a sequence of local measures μ_t (Remark A.2). With a line search to get properly large μ_t, this could be helpful in high dimensional sparse learning problems which may or may *not* have restricted strong convexity (the associated parameter often hard to determine in theory).

5. Summary. This paper studied the class of iterative algorithms derived from GBF-defined surrogates with a possible lack of convexity and/or smoothness. These surrogates differ from the MM surrogates frequently used in statistical computation, in that they gain additional first-order degeneracy and may drop the majorization requirement. GBFs have interesting connections to the densities in the exponential family and possess some idempotence properties that are useful for studying iterative algorithms.

The GBF calculus built by the lemmas not only facilitates optimization error analysis but can be bound to the empirical process theory for nonasymptotic statistical analysis (cf. Sections 3.2 and A.18). In addition to obtaining some insightful results in the realm of convex optimization, we were able to build universal global convergence rates for a broad class of

Bregman-surrogate algorithms for nonsmooth nonconvex optimization. Moreover, in the nonconvex composite setting that is of great interest in high dimensional statistics, we found that the sequence of iterates generated by Bregman surrogates can approach the statistical truth at a linear rate even when $p > n$, and the obtained fixed points enjoy oracle inequalities with essentially the optimal order of statistical accuracy, under some regularity conditions less demanding than those used in the literature. Finally, we devised two "all-in-one" acceleration schemes with novel updates of the iterates and relaxation and stepsize parameters, and some sharp theoretical bounds were shown without assuming smoothness or convexity.

APPENDIX A: PROOFS

We list some notation and symbols that are used in the proofs. Given a directionally differentiable function ψ, $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \delta\psi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})$, $\bar{\boldsymbol{\Delta}}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = (\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\gamma},\boldsymbol{\beta}))/2$, and $\grave{\boldsymbol{\Delta}}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\boldsymbol{\gamma},\boldsymbol{\beta})$. We occasionally denote $\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma})$ by $\boldsymbol{\Delta}(\boldsymbol{\beta},\boldsymbol{\gamma})$ when there is no ambiguity. The classes of continuous functions and continuously differentiable functions are denoted by $\mathcal{C}^0$ and $\mathcal{C}^1$, respectively. Recall that all functions are assumed to be defined on a vector space unless otherwise mentioned.

DEFINITION A.1. *We call ξ a sub-Gaussian random variable if and only if there exist constants $C, c > 0$ such that $\mathbb{P}\{|\xi| \geq t\} \leq Ce^{-ct^2}, \forall t > 0$. The scale (or ψ_2-norm) of ξ is defined by $\sigma(\xi) = \inf\{\sigma > 0 : \mathbb{E}\exp(\xi^2/\sigma^2) \leq 2\}$. More generally, $\boldsymbol{\xi} \in \mathbb{R}^p$ is called a sub-Gaussian random vector with scale bounded by σ if all one-dimensional marginals $\langle\boldsymbol{\xi},\boldsymbol{\alpha}\rangle$ are sub-Gaussian satisfying $\|\langle\boldsymbol{\xi},\boldsymbol{\alpha}\rangle\|_{\psi_2} \leq \sigma\|\boldsymbol{\alpha}\|_2$, $\forall\boldsymbol{\alpha} \in \mathbb{R}^p$.*

DEFINITION A.2. *We call d a pseudo-metric if it satisfies $d(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) = d(\boldsymbol{\eta}_2,\boldsymbol{\eta}_1) \geq 0$ and $d(\boldsymbol{\eta}_1,\boldsymbol{\eta}_2) \leq d(\boldsymbol{\eta}_1,\boldsymbol{\eta}_3) + d(\boldsymbol{\eta}_2,\boldsymbol{\eta}_3)$, for all $\boldsymbol{\eta}_1,\boldsymbol{\eta}_2,\boldsymbol{\eta}_3$.*

We state a first-order optimality condition satisfied by all local minimizers of f that is directionally differentiable. The result is basic and we omit the proof. It holds the key to deriving the so-called "basic inequality" in a variety of statistical learning problems.

LEMMA A.1. *Let $f : \mathbb{R}^p \to \mathbb{R}$ be a real-valued function and $C \subset \mathbb{R}^p$ be a convex set. Suppose that f is directionally differentiable at $\boldsymbol{\beta}^o$ that is a local minimizer to the problem $\min_{\boldsymbol{\beta}\in C} f(\boldsymbol{\beta})$. Then $\delta f(\boldsymbol{\beta}^o;\boldsymbol{h}) \geq 0$ with $\boldsymbol{h} = \boldsymbol{\beta} - \boldsymbol{\beta}^o$ or $f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^o) \geq \boldsymbol{\Delta}_f(\boldsymbol{\beta},\boldsymbol{\beta}^o)$ for all $\boldsymbol{\beta} \in C$.*

A.1. Proof of Lemma 1. (i) This property is straightforward by definition:

$$\begin{aligned}
&\boldsymbol{\Delta}_{a\psi+b\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) \\
&= (a\psi + b\varphi)(\boldsymbol{\beta}) - (a\psi + b\varphi)(\boldsymbol{\gamma}) - \delta(a\psi + b\varphi)(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma}) \\
&= a\big[\psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \delta\psi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})\big] + b\big[\varphi(\boldsymbol{\beta}) - \varphi(\boldsymbol{\gamma}) - \delta\varphi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})\big] \\
&= a\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) + b\boldsymbol{\Delta}_\varphi(\boldsymbol{\beta},\boldsymbol{\gamma}).
\end{aligned}$$

(ii) From [45, Theorem 23.1], the convexity of ψ implies the directional differentiability of ψ and the positively homogenous convexity of $\delta\psi(\boldsymbol{\beta};\cdot)$ for any given $\boldsymbol{\beta}$, and we can write

$$\delta\psi(\boldsymbol{\beta};\boldsymbol{h}) = \inf_{\epsilon>0} \frac{\psi(\boldsymbol{\beta}+\epsilon\boldsymbol{h}) - \psi(\boldsymbol{\beta})}{\epsilon}. \tag{A.1}$$

Putting $\epsilon = 1$ and $\boldsymbol{h} = \boldsymbol{\gamma} - \boldsymbol{\beta}$ in (A.1) gives $\delta\psi(\boldsymbol{\beta};\boldsymbol{\gamma}-\boldsymbol{\beta}) \leq \psi(\boldsymbol{\gamma}) - \psi(\boldsymbol{\beta})$, thus $\boldsymbol{\Delta}_\psi(\boldsymbol{\gamma},\boldsymbol{\beta}) \geq 0$.

Conversely, suppose that ψ defined on $\mathbb{R}^n$ is directionally differentiable ($\delta\psi$ exists and is finite), thus radially continuous, and $\boldsymbol{\Delta}_\psi \geq 0$. For any $\boldsymbol{s}_{\boldsymbol{\beta}} : \langle \boldsymbol{s}_{\boldsymbol{\beta}}, \boldsymbol{h}\rangle \leq \delta\psi(\boldsymbol{\beta};\boldsymbol{h}), \boldsymbol{s}_{\boldsymbol{\gamma}} : \langle \boldsymbol{s}_{\boldsymbol{\gamma}}, \boldsymbol{h}\rangle \leq \delta\psi(\boldsymbol{\gamma};\boldsymbol{h})\ \forall \boldsymbol{h}$,

$$\psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) - \langle \boldsymbol{s}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}-\boldsymbol{\gamma}\rangle \geq \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) \geq 0, \tag{A.2}$$

$$\psi(\boldsymbol{\gamma}) - \psi(\boldsymbol{\beta}) - \langle \boldsymbol{s}_{\boldsymbol{\beta}}, \boldsymbol{\gamma}-\boldsymbol{\beta}\rangle \geq \boldsymbol{\Delta}_\psi(\boldsymbol{\gamma},\boldsymbol{\beta}) \geq 0. \tag{A.3}$$

Adding them together gives $\langle \boldsymbol{s}_{\boldsymbol{\beta}} - \boldsymbol{s}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}-\boldsymbol{\gamma}\rangle \geq 0$. This indicates the monotone property of the Clarke-Rockafellar subdifferential of ψ, thereby its convexity according to [16].

(iii) To show the first result, notice that $\boldsymbol{\Delta}_{\psi\circ\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) - \boldsymbol{\Delta}_\psi(\varphi(\boldsymbol{\beta}),\varphi(\boldsymbol{\gamma})) = \lim_{\epsilon\to 0+}\{\psi(\varphi(\boldsymbol{\gamma}) + \epsilon(\varphi(\boldsymbol{\beta})-\varphi(\boldsymbol{\gamma}))) - \psi(\varphi(\boldsymbol{\gamma}+\epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma})))\}/\epsilon = \langle \nabla\psi(\varphi(\boldsymbol{\gamma})), \varphi(\boldsymbol{\beta})-\varphi(\boldsymbol{\gamma})\rangle - \delta(\psi\circ\varphi)(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})$. From $\psi \in \mathcal{C}^1$ and $\varphi \in \mathcal{C}^0$,

$$\begin{aligned}
\delta(\psi\circ\varphi)(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma}) &= \lim_{\epsilon\to 0+}\{\psi(\varphi(\boldsymbol{\gamma}) + (\varphi(\boldsymbol{\gamma}+\epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma})) - \varphi(\boldsymbol{\gamma}))) - \psi(\varphi(\boldsymbol{\gamma}))\}/\epsilon \\
&= \lim_{\epsilon\to 0+}\langle \nabla\psi(\varphi(\boldsymbol{\gamma})), \varphi(\boldsymbol{\gamma}+\epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma})) - \varphi(\boldsymbol{\gamma})/\epsilon\rangle \\
&= \langle \nabla\psi(\varphi(\boldsymbol{\gamma})), \delta\varphi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})\rangle.
\end{aligned}$$

Using the definition of $\boldsymbol{\Delta}_\varphi(\boldsymbol{\beta},\boldsymbol{\gamma})$ (the componentwise extension), we obtain the conclusion.

Next, we prove the second result. Let $\varphi : \mathbb{R}^p \to \mathbb{R}^n$ be the linear function $\varphi(\boldsymbol{\beta}) = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\alpha}$ with its Jacobian matrix $\mathrm{D}\varphi(\boldsymbol{\beta}) := [\mathrm{D}_j\varphi_i(x)] = \boldsymbol{X} \in \mathbb{R}^{n\times p}$. By definition, $\boldsymbol{\Delta}_{\psi\circ\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \psi(\varphi(\boldsymbol{\beta})) - \psi(\varphi(\boldsymbol{\gamma})) - \delta(\psi\circ\varphi)(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})$ and

$$\begin{aligned}
\delta(\psi\circ\varphi)(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma}) &= \lim_{\epsilon\to 0+}\{\psi(\varphi(\boldsymbol{\gamma}+\epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma}))) - \psi(\varphi(\boldsymbol{\gamma}))\}/\epsilon \\
&= \lim_{\epsilon\to 0+}\{\psi(\varphi(\boldsymbol{\gamma}) + \epsilon\,\mathrm{D}\varphi(\boldsymbol{\gamma})(\boldsymbol{\beta}-\boldsymbol{\gamma}))) - \psi(\varphi(\boldsymbol{\gamma}))\}/\epsilon \\
&= \delta\psi(\varphi(\boldsymbol{\gamma});\, \mathrm{D}\varphi(\boldsymbol{\gamma})(\boldsymbol{\beta}-\boldsymbol{\gamma})) = \delta\psi(\varphi(\boldsymbol{\gamma});\varphi(\boldsymbol{\beta}) - \varphi(\boldsymbol{\gamma})),
\end{aligned}$$

from which it follows that $\boldsymbol{\Delta}_{\psi\circ\varphi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \boldsymbol{\Delta}_\psi(\varphi(\boldsymbol{\beta}),\varphi(\boldsymbol{\gamma}))$.

(iv) From Theorem 11 in [25], for any continuous function f with finite Dini derivative $D^+f(x) := \limsup_{\epsilon\to 0+}(f(x+\epsilon) - f(x))/\epsilon$, if $D^+f(x)$ is integrable over $[a,b]$, $f(b) - f(a) = \int_a^b D^+f(x)\mathrm{d}x$. By definition, ψ is continuous when restricted to the line segment $[\boldsymbol{\beta},\boldsymbol{\gamma}]$ (radial continuity). It follows that

$$\begin{aligned}
\psi(\boldsymbol{\beta}) - \psi(\boldsymbol{\gamma}) &= \psi\big(\boldsymbol{\gamma} + t(\boldsymbol{\beta}-\boldsymbol{\gamma})\big)\Big|_{t=0}^1 \\
&= \int_0^1 \lim_{\epsilon\to 0+}\frac{1}{\epsilon}\Big[\psi\big(\boldsymbol{\gamma} + (t+\epsilon)(\boldsymbol{\beta}-\boldsymbol{\gamma})\big) - \psi\big(\boldsymbol{\gamma} + t(\boldsymbol{\beta}-\boldsymbol{\gamma})\big)\Big]\mathrm{d}t \\
&= \int_0^1 \lim_{\epsilon\to 0+}\frac{1}{\epsilon}\Big[\psi\big(\boldsymbol{\gamma} + t(\boldsymbol{\beta}-\boldsymbol{\gamma}) + \epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma})\big) - \psi\big(\boldsymbol{\gamma} + t(\boldsymbol{\beta}-\boldsymbol{\gamma})\big)\Big]\mathrm{d}t \\
&= \int_0^1 \delta\psi\big(\boldsymbol{\gamma} + t(\boldsymbol{\beta}-\boldsymbol{\gamma});\, \boldsymbol{\beta}-\boldsymbol{\gamma}\big)\mathrm{d}t.
\end{aligned}$$

Hence, $\boldsymbol{\Delta}_\psi$ can be formulated by

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \int_0^1 \Big[\delta\psi\big(\boldsymbol{\gamma} + t(\boldsymbol{\beta}-\boldsymbol{\gamma});\, \boldsymbol{\beta}-\boldsymbol{\gamma}\big) - \delta\psi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})\Big]\,\mathrm{d}t.$$

A.2. Proof of Lemma 2. (i) First, if $\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})$ is directionally differentiable, then

$$\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma})-\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})=\boldsymbol{\Delta}_{\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) \tag{A.4}$$

for any $\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma}$. In fact, $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma})-\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})=\boldsymbol{\Delta}_{\psi(\cdot)-\boldsymbol{\Delta}_{\psi}(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})=\boldsymbol{\Delta}_{\psi(\boldsymbol{\alpha})+\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})=\boldsymbol{\Delta}_{\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})$.

Accordingly, when ψ is convex, which means $\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})$ is convex as well (cf. Section A.1), $\boldsymbol{\Delta}_{\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})\geq 0$ by Lemma 1. The result under concavity can be similarly proved.

(ii) Let

$$q(\cdot;\boldsymbol{\alpha})=\delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha}). \tag{A.5}$$

We want to show for $\boldsymbol{\alpha}=\theta\boldsymbol{\beta}+(1-\theta)\boldsymbol{\gamma}$ with $\theta\leq 0$ or $\theta\geq 1$, $\boldsymbol{\Delta}_{q(\cdot;\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})$ is well-defined and equals 0. This is intuitive due to the linearity of q when restricted to $[\boldsymbol{\beta},\boldsymbol{\gamma}]$, assuming $\boldsymbol{\beta}-\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}-\boldsymbol{\alpha}$ are positively collinear.

To verify it, by definition,

$$\begin{aligned}\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})&=\lim_{\epsilon\to 0+}[q(\boldsymbol{\gamma}+\epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma});\boldsymbol{\alpha})-q(\boldsymbol{\gamma};\boldsymbol{\alpha})]/\epsilon\\&=\lim_{\epsilon\to 0+}[\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}+\epsilon(\boldsymbol{\beta}-\boldsymbol{\gamma})-\boldsymbol{\alpha})-\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\alpha})]/\epsilon\\&=\lim_{\epsilon\to 0+}[\delta\psi(\boldsymbol{\alpha};(\theta-\epsilon)(\boldsymbol{\gamma}-\boldsymbol{\beta}))-\delta\psi(\boldsymbol{\alpha};\theta(\boldsymbol{\gamma}-\boldsymbol{\beta}))]/\epsilon\\&=\lim_{\epsilon\to 0+}[\delta\psi(\boldsymbol{\alpha};(\epsilon-\theta)(\boldsymbol{\beta}-\boldsymbol{\gamma}))-\delta\psi(\boldsymbol{\alpha};(-\theta)(\boldsymbol{\beta}-\boldsymbol{\gamma}))]/\epsilon,\end{aligned}$$

and so with $\theta>0$,

$$\begin{aligned}\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})&=\lim_{\epsilon\to 0+}[(\theta-\epsilon)\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\beta})-\theta\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\beta})]/\epsilon\\&=-\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\beta}),\end{aligned}$$

and with $\theta\leq 0$,

$$\begin{aligned}\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})&=\lim_{\epsilon\to 0+}[(\epsilon-\theta)\delta\psi(\boldsymbol{\alpha};(\boldsymbol{\beta}-\boldsymbol{\gamma}))-(-\theta)\delta\psi(\boldsymbol{\alpha};(\boldsymbol{\beta}-\boldsymbol{\gamma}))]/\epsilon\\&=\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\gamma}).\end{aligned}$$

The above derivation also guarantees the existence of $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})$. Now, as $\theta\geq 1$, $\langle\boldsymbol{\beta}-\boldsymbol{\alpha},\boldsymbol{\gamma}-\boldsymbol{\beta}\rangle\geq 0$ and so $q(\boldsymbol{\beta};\boldsymbol{\alpha})-q(\boldsymbol{\gamma};\boldsymbol{\alpha})-\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})=\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\alpha})-\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\alpha})+\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\beta})=0$. As $\theta\leq 0$, $\langle\boldsymbol{\beta}-\boldsymbol{\alpha},\boldsymbol{\gamma}-\boldsymbol{\beta}\rangle\leq 0$ and $q(\boldsymbol{\beta};\boldsymbol{\alpha})-q(\boldsymbol{\gamma};\boldsymbol{\alpha})-\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})=\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\alpha})-\delta\psi(\boldsymbol{\alpha};\boldsymbol{\gamma}-\boldsymbol{\alpha})-\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\gamma})=0$.

(iii) By definition, we have

$$\begin{aligned}&\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{z};\boldsymbol{\beta}-\boldsymbol{\gamma})\\&=\lim_{\epsilon_2\to 0+}\frac{1}{\epsilon_2}\{q(\boldsymbol{z}+\epsilon_2(\boldsymbol{\beta}-\boldsymbol{\gamma});\boldsymbol{\alpha})-q(\boldsymbol{z};\boldsymbol{\alpha})\}\\&=\lim_{\epsilon_2\to 0+}\frac{1}{\epsilon_2}\{\delta\psi(\boldsymbol{\alpha};\boldsymbol{z}+\epsilon_2(\boldsymbol{\beta}-\boldsymbol{\gamma})-\boldsymbol{\alpha})-\delta\psi(\boldsymbol{\alpha};\boldsymbol{z}-\boldsymbol{\alpha})\}.\end{aligned}$$

Under the restricted linearity condition $\delta\psi(\boldsymbol{\alpha};\boldsymbol{h})=\langle g(\boldsymbol{\alpha}),\boldsymbol{h}\rangle,\forall\boldsymbol{h}\in[\boldsymbol{\beta}-\boldsymbol{\alpha},\boldsymbol{\gamma}-\boldsymbol{\alpha}]$, for $\boldsymbol{z}=\boldsymbol{\gamma}+t(\boldsymbol{\beta}-\boldsymbol{\gamma})$ with $t\in[0,1)$,

$$\begin{aligned}\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{z};\boldsymbol{\beta}-\boldsymbol{\gamma})&=\lim_{\epsilon_2\to 0+}\frac{1}{\epsilon_2}\langle g(\boldsymbol{\alpha}),\boldsymbol{z}+\epsilon_2(\boldsymbol{\beta}-\boldsymbol{\gamma})-\boldsymbol{\alpha}-\boldsymbol{z}+\boldsymbol{\alpha}\rangle\\&=\langle g(\boldsymbol{\alpha}),\boldsymbol{\beta}-\boldsymbol{\gamma}\rangle.\end{aligned}$$

Under the restricted continuity condition $\lim_{\epsilon\to 0+}\delta\psi(\boldsymbol{\alpha}+\epsilon\boldsymbol{h};\boldsymbol{\beta}-\boldsymbol{\gamma})=\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\gamma}),\forall\boldsymbol{h}\in[\boldsymbol{\beta}-\boldsymbol{\alpha},\boldsymbol{\gamma}-\boldsymbol{\alpha}]$, for $\boldsymbol{z}=\boldsymbol{\gamma}+t(\boldsymbol{\beta}-\boldsymbol{\gamma})$ with $t\in[0,1)$,

$$
\begin{aligned}
&\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{z};\boldsymbol{\beta}-\boldsymbol{\gamma})\\
&=\lim_{\epsilon_2\to 0+}\frac{1}{\epsilon_2}\Bigg\{\lim_{\epsilon_1\to 0+}\frac{1}{\epsilon_1}\Big[\psi\big(\boldsymbol{\alpha}+\epsilon_1[\boldsymbol{z}+\epsilon_2(\boldsymbol{\beta}-\boldsymbol{\gamma})-\boldsymbol{\alpha}]\big)-\psi(\boldsymbol{\alpha})\Big]\\
&\qquad\qquad-\lim_{\epsilon_1\to 0+}\frac{1}{\epsilon_1}\Big[\psi\big(\boldsymbol{\alpha}+\epsilon_1(\boldsymbol{z}-\boldsymbol{\alpha})\big)-\psi(\boldsymbol{\alpha})\Big]\Bigg\}\\
&=\lim_{\epsilon_2\to 0+}\lim_{\epsilon_1\to 0+}\frac{1}{\epsilon_1\epsilon_2}\Big[\psi\big((1-\epsilon_1)\boldsymbol{\alpha}+\epsilon_1\boldsymbol{z}-\epsilon_1\epsilon_2\boldsymbol{\gamma}+\epsilon_1\epsilon_2\boldsymbol{\beta}\big)-\psi\big((1-\epsilon_1)\boldsymbol{\alpha}+\epsilon_1\boldsymbol{z}\big)\Big]\\
&=\lim_{\epsilon_2\to 0+}\lim_{\epsilon_1\to 0+}\frac{1}{\epsilon_1\epsilon_2}\int_0^1\delta\psi\big((1-\epsilon_1)\boldsymbol{\alpha}+\epsilon_1\boldsymbol{z}+\epsilon_1\epsilon_2 s(\boldsymbol{\beta}-\boldsymbol{\gamma});\epsilon_1\epsilon_2(\boldsymbol{\beta}-\boldsymbol{\gamma})\big)\,\mathrm{d}s\\
&=\lim_{\epsilon_2\to 0+}\lim_{\epsilon_1\to 0+}\int_0^1\delta\psi\big((1-\epsilon_1)\boldsymbol{\alpha}+\epsilon_1\boldsymbol{z}+\epsilon_1\epsilon_2 s(\boldsymbol{\beta}-\boldsymbol{\gamma});\boldsymbol{\beta}-\boldsymbol{\gamma}\big)\,\mathrm{d}s\\
&=\lim_{\epsilon_2\to 0+}\int_0^1\lim_{\epsilon_1\to 0+}\delta\psi\big(\boldsymbol{\alpha}+\epsilon_1(\boldsymbol{z}+\epsilon_2 s(\boldsymbol{\beta}-\boldsymbol{\gamma})-\boldsymbol{\alpha});\boldsymbol{\beta}-\boldsymbol{\gamma}\big)\,\mathrm{d}s\\
&=\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\gamma}),
\end{aligned}
$$

where we used the positive homogeneity of $\delta\psi(\boldsymbol{\alpha};\cdot)$ and the dominated convergence theorem. (The integral is well-defined due to the boundedness and Lebesgue measurability of the integrand.)

The two sets of conditions are not equivalent in multiple dimensions. But in either case, $\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{z};\boldsymbol{\beta}-\boldsymbol{\gamma})$ is a term independent of $\boldsymbol{z}$. Hence by Lemma 1 (iv),

$$
\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma})=\int_0^1\Big[\delta q(\cdot;\boldsymbol{\alpha})\big(\boldsymbol{\gamma}+t(\boldsymbol{\beta}-\boldsymbol{\gamma});\,\boldsymbol{\beta}-\boldsymbol{\gamma}\big)-\delta q(\cdot;\boldsymbol{\alpha})(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})\Big]\mathrm{d}t=0.
$$

A.3. Proof of Lemma 3. (i) Let $\varphi=b^*$. Then for all subgradient $\boldsymbol{g}\in\partial\varphi(\boldsymbol{z})$, $(\boldsymbol{g},\boldsymbol{z})$ makes a conjugate pair and so $\langle\boldsymbol{g},\boldsymbol{z}\rangle=b(\boldsymbol{g})+\varphi(\boldsymbol{z})$ (see, e.g., [45]). Using the shorthand notation $(\partial\varphi(\boldsymbol{z}),\boldsymbol{z})$, we represent it as $\langle\partial\varphi(\boldsymbol{z}),\boldsymbol{z}\rangle=b(\partial\varphi(\boldsymbol{z}))+\varphi(\boldsymbol{z})$. Therefore,

$$
\begin{aligned}
\sigma^2 l_0(\boldsymbol{\eta};\boldsymbol{z})+b^*(\boldsymbol{z})&=-\langle\boldsymbol{z},\boldsymbol{\eta}\rangle+b(\boldsymbol{\eta})+\varphi(\boldsymbol{z})\\
&=-\langle\boldsymbol{z},\boldsymbol{\eta}\rangle+b(\boldsymbol{\eta})+\langle\boldsymbol{z},\partial\varphi(\boldsymbol{z})\rangle-b(\partial\varphi(\boldsymbol{z}))\\
&=b(\boldsymbol{\eta})-b(\partial\varphi(\boldsymbol{z}))-\langle\boldsymbol{z},\boldsymbol{\eta}-\partial\varphi(\boldsymbol{z})\rangle\\
&=b(\boldsymbol{\eta})-b(\partial\varphi(\boldsymbol{z}))-\langle\nabla b(\partial\varphi(\boldsymbol{z})),\boldsymbol{\eta}-\partial\varphi(\boldsymbol{z})\rangle\\
&=\boldsymbol{\Delta}_b(\boldsymbol{\eta},\partial\varphi(\boldsymbol{z})).
\end{aligned}
$$

When $p_{\boldsymbol{\eta}}$ is minimal, $\mathcal{M}$ is full-dimensional and the canonical link $g=(\nabla b)^{-1}$ is well-defined on $\mathcal{M}^\circ$ (Proposition 3.1 and Proposition 3.2 in [62] can be slightly modified to include the dispersion parameter), and so $(g(\boldsymbol{z}),\nabla b(g(\boldsymbol{z})))$ or $(g(\boldsymbol{z}),\boldsymbol{z})$ makes a conjugate pair.

(ii) Let $\boldsymbol{\mu}(\boldsymbol{\eta})=\nabla b(\boldsymbol{\eta})$ or $\boldsymbol{\mu}$ for brevity. It follows that $\boldsymbol{\eta}\in\partial\varphi(\boldsymbol{\mu})$ and so

$$
\begin{aligned}
-\langle\boldsymbol{z},\boldsymbol{\eta}\rangle+b(\boldsymbol{\eta})+b^*(\boldsymbol{z})&=-\langle\boldsymbol{z},\boldsymbol{\eta}\rangle+\langle\boldsymbol{\mu},\boldsymbol{\eta}\rangle-\varphi(\boldsymbol{\mu})+\varphi(\boldsymbol{z})\\
&=-\langle\boldsymbol{z}-\boldsymbol{\mu},\boldsymbol{\eta}\rangle-\varphi(\boldsymbol{\mu})+\varphi(\boldsymbol{z})\\
&\geq-\delta\varphi(\boldsymbol{\mu};\boldsymbol{z}-\boldsymbol{\mu})-\varphi(\boldsymbol{\mu})+\varphi(\boldsymbol{z})=\boldsymbol{\Delta}_\varphi(\boldsymbol{z},\boldsymbol{\mu}),
\end{aligned}
$$

where the inequality is due to [45, Theorem 23.2]. We claim that the inequality is actually an equality.

Indeed, if there exist $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2 \in \partial\varphi(\boldsymbol{\mu})$ with $\boldsymbol{\eta}_1 \neq \boldsymbol{\eta}_2$, then $\bar{\boldsymbol{\Delta}}_b(\boldsymbol{\eta}_2, \boldsymbol{\eta}_1) = \langle \nabla b(\boldsymbol{\eta}_2) - \nabla b(\boldsymbol{\eta}_1), \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1 \rangle = 0$ and so $\boldsymbol{\Delta}_b(\boldsymbol{\eta}_2, \boldsymbol{\eta}_1) = 0$ since b is convex. Therefore, for any random vector $\boldsymbol{y}$ following $p_{\boldsymbol{\eta}}$ in the exponential family, where $\boldsymbol{\eta} = t\boldsymbol{\eta}_1 + (1-t)\boldsymbol{\eta}_2$, $t \in (0,1)$,

$$Var((\boldsymbol{\eta}_2 - \boldsymbol{\eta}_1)^T \boldsymbol{y}) = 0,$$

which can be obtained from Proposition 3.1 of [62]. Because $\exp((\langle \cdot, \boldsymbol{\eta} \rangle - b(\boldsymbol{\eta}))/\sigma^2) > 0$ for any $\boldsymbol{\eta} \in \Omega$, we have $\langle \boldsymbol{\eta}_2 - \boldsymbol{\eta}_1, \boldsymbol{z} \rangle = c$ for almost every $\boldsymbol{z} \in \mathcal{Y}^n$ with respect to ν (i.e., $p_{\boldsymbol{\eta}}$ is not minimal). It follows that

$$\langle \boldsymbol{z} - \boldsymbol{\mu}, \boldsymbol{\eta}_1 - \boldsymbol{\eta}_2 \rangle = 0.$$

Finally, from $\delta\varphi(\boldsymbol{\mu}; \boldsymbol{h}) = \sup\{\langle \boldsymbol{g}, \boldsymbol{h} \rangle : \boldsymbol{g} \in \partial\varphi(\boldsymbol{\mu})\}$ [45, Theorem 23.4], the claim is true.

In the case that $p_{\boldsymbol{\eta}}$ is also minimal, φ can be shown to be strictly convex and differentiable on $\mathcal{M}^\circ$ [45, Theorem 26.4].

(iii) Let $\mathrm{d}P_{\boldsymbol{\eta}} = p_{\boldsymbol{\eta}} \, \mathrm{d}\nu_0$. By definition,

$$\mathrm{KL}(p_{\boldsymbol{\eta}_1}, p_{\boldsymbol{\eta}_2}) = \int \log(\mathrm{d}P_{\boldsymbol{\eta}_1} / \mathrm{d}P_{\boldsymbol{\eta}_2}) \, \mathrm{d}P_{\boldsymbol{\eta}_1} = \int p_{\boldsymbol{\eta}_1} \log(p_{\boldsymbol{\eta}_1} / \mathrm{d}p_{\boldsymbol{\eta}_2}) \, \mathrm{d}\nu_0$$

and so

$$\begin{aligned}
\mathrm{KL}(p_{\boldsymbol{\eta}_1}, p_{\boldsymbol{\eta}_2}) &= \int \log \{ e^{(\langle \boldsymbol{y}, \boldsymbol{\eta}_1 \rangle - b(\boldsymbol{\eta}_1))/\sigma^2 - c(\boldsymbol{y}, \sigma^2)} / e^{(\langle \boldsymbol{y}, \boldsymbol{\eta}_2 \rangle - b(\boldsymbol{\eta}_2))/\sigma^2 - c(\boldsymbol{y}, \sigma^2)} \} \, \mathrm{d}P_{\boldsymbol{\eta}_1} \\
&= \frac{1}{\sigma^2} \int \langle \boldsymbol{y}, \boldsymbol{\eta}_1 - \boldsymbol{\eta}_2 \rangle - b(\boldsymbol{\eta}_1) + b(\boldsymbol{\eta}_2) \, \mathrm{d}P_{\boldsymbol{\eta}_1} \\
&= \frac{1}{\sigma^2} \{ b(\boldsymbol{\eta}_2) - b(\boldsymbol{\eta}_1) + \int \langle \boldsymbol{y}, \boldsymbol{\eta}_1 - \boldsymbol{\eta}_2 \rangle \, \mathrm{d}P_{\eta_1} \} \\
&= \frac{1}{\sigma^2} \{ b(\boldsymbol{\eta}_2) - b(\boldsymbol{\eta}_1) + \langle \nabla b(\boldsymbol{\eta}_1), \boldsymbol{\eta}_1 - \boldsymbol{\eta}_2 \rangle \} \\
&= \boldsymbol{\Delta}_b(\boldsymbol{\eta}_2, \boldsymbol{\eta}_1)/\sigma^2,
\end{aligned}$$

where the third equality is due to $\mathbb{E}_{\boldsymbol{y} \sim p_{\boldsymbol{\eta}_1}} \boldsymbol{y} = \nabla b(\boldsymbol{\eta}_1)$ under $\boldsymbol{\eta}_1 \in \Omega^\circ$ (which can be derived from Proposition 3.1 of [62]). Moreover, from Lemma 1, $\sigma^2 \boldsymbol{\Delta}_{l_0}(\boldsymbol{\eta}_2, \boldsymbol{\eta}_1) = \boldsymbol{\Delta}_b(\boldsymbol{\eta}_2, \boldsymbol{\eta}_1)$.

A.4. Proof of Lemma 4. In this proof, all directional derivatives are with respect with $\boldsymbol{\beta}$. The result of (i) is trivial from the construction of g. For (ii), by definition, we have $\delta g(\boldsymbol{\beta}; \boldsymbol{\beta}^-, \boldsymbol{h}) = \delta f(\boldsymbol{\beta}; \boldsymbol{h}) + \delta\psi(\boldsymbol{\beta}; \boldsymbol{h}) - \delta q(\boldsymbol{\beta}; \boldsymbol{\beta}^-, \boldsymbol{h})$ with $q(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = \delta\psi(\boldsymbol{\beta}^-; \boldsymbol{\beta} - \boldsymbol{\beta}^-)$. It follows from $q(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = \lim_{\epsilon \to 0+} [\psi(\boldsymbol{\beta}^- + \epsilon(\boldsymbol{\beta} - \boldsymbol{\beta}^-)) - \psi(\boldsymbol{\beta}^-)]/\epsilon$ that

$$\begin{aligned}
\delta q(\boldsymbol{\beta}; \boldsymbol{\beta}^-, \boldsymbol{h}) &= \lim_{\epsilon' \to 0+} [q(\boldsymbol{\beta} + \epsilon' \boldsymbol{h}; \boldsymbol{\beta}^-) - q(\boldsymbol{\beta}; \boldsymbol{\beta}^-)]/\epsilon' \\
&= \lim_{\epsilon' \to 0+} \big\{ (1/\epsilon') \lim_{\epsilon \to 0+} [\psi(\boldsymbol{\beta}^- + \epsilon(\boldsymbol{\beta} + \epsilon' \boldsymbol{h} - \boldsymbol{\beta}^-)) - \psi(\boldsymbol{\beta}^-)]/\epsilon \big\} \\
&\quad - \lim_{\epsilon' \to 0+} \delta\psi(\boldsymbol{\beta}^-; \boldsymbol{\beta} - \boldsymbol{\beta}^-)/\epsilon'.
\end{aligned}$$

When $\boldsymbol{\beta}^- = \boldsymbol{\beta}$, $\delta\psi(\boldsymbol{\beta}^-; \boldsymbol{\beta} - \boldsymbol{\beta}^-) = 0$ and so

$$\begin{aligned}
\delta q(\boldsymbol{\beta}; \boldsymbol{\beta}^-, \boldsymbol{h})|_{\boldsymbol{\beta}^- = \boldsymbol{\beta}} &= \lim_{\epsilon' \to 0+} \lim_{\epsilon \to 0+} [\psi(\boldsymbol{\beta} + \epsilon(\epsilon' \boldsymbol{h})) - \psi(\boldsymbol{\beta})]/(\epsilon\epsilon') \\
&= \lim_{\epsilon'' \to 0+} [\psi(\boldsymbol{\beta} + \epsilon'' \boldsymbol{h}) - \psi(\boldsymbol{\beta})]/\epsilon'' \\
&= \delta\psi(\boldsymbol{\beta}; \boldsymbol{h}).
\end{aligned}$$

The above argument also guarantees the existence of $\delta g(\boldsymbol{\beta};\boldsymbol{\beta}^-,\boldsymbol{h})|_{\boldsymbol{\beta}^-=\boldsymbol{\beta}}$. Therefore, $\delta g(\boldsymbol{\beta};\boldsymbol{\beta}^-,\boldsymbol{h})|_{\boldsymbol{\beta}^-=\boldsymbol{\beta}} = \delta f(\boldsymbol{\beta};\boldsymbol{h})$ for any $\boldsymbol{\beta}$ and $\boldsymbol{h}$.

A.5. Proof of Lemma 5. All results in Lemma 1 and Lemma 2 can be formulated for $\mathbf{C}$. For example, ψ is convex if and only if $\mathbf{C}_\psi \geq 0$, $\mathbf{C}_{a\phi+b\varphi} = a\mathbf{C}_\phi + b\mathbf{C}_\varphi$, $\boldsymbol{\Delta}_\psi \geq \mu\mathbf{D}_2$ implies $\mathbf{C}_\psi \geq \mu\mathbf{C}_2$ since $\mathbf{C}_2(\boldsymbol{\alpha},\boldsymbol{\beta},\theta) := \mathbf{C}_{\|\cdot\|_2^2/2}(\boldsymbol{\alpha},\boldsymbol{\beta},\theta) = \theta(1-\theta)\mathbf{D}_2(\boldsymbol{\alpha},\boldsymbol{\beta})$, and so on. To show (i), we have

$$\begin{aligned}
&\mathbf{C}_\psi(\boldsymbol{\alpha},\boldsymbol{\beta},\theta) + \boldsymbol{\Delta}_\psi(\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta},\boldsymbol{\alpha}) \\
=&\theta\psi(\boldsymbol{\alpha}) + (1-\theta)\psi(\boldsymbol{\beta}) - \psi(\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta}) \\
&+ \psi(\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta}) - \psi(\boldsymbol{\alpha}) - \delta\psi(\boldsymbol{\alpha};\theta\boldsymbol{\alpha}+(1-\theta)\boldsymbol{\beta}-\boldsymbol{\alpha}) \\
=&(\theta-1)\psi(\boldsymbol{\alpha}) + (1-\theta)\psi(\boldsymbol{\beta}) - \delta\psi(\boldsymbol{\alpha};(1-\theta)(\boldsymbol{\beta}-\boldsymbol{\alpha})) \\
=&(1-\theta)\psi(\boldsymbol{\beta}) - (1-\theta)\psi(\boldsymbol{\alpha}) - (1-\theta)\delta\psi(\boldsymbol{\alpha};\boldsymbol{\beta}-\boldsymbol{\alpha}) \\
=&(1-\theta)\boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\alpha}).
\end{aligned}$$

Similar to the proof of Lemma 2, let $q(\cdot;\boldsymbol{\alpha}) = \delta\psi(\boldsymbol{\alpha};\cdot-\boldsymbol{\alpha})$. Then

$$\mathbf{C}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma},\theta) - \mathbf{C}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma},\theta) = \mathbf{C}_{\psi(\boldsymbol{\alpha})+q(\cdot;\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma},\theta) = \mathbf{C}_{q(\cdot;\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma},\theta),$$

without requiring the directional differentiability of $q(\cdot;\boldsymbol{\alpha})$. We can show analogous results to Lemma 2. For example, for any convex ψ, from the positively homogenous convexity of q,

$$\mathbf{C}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})} \leq \mathbf{C}_\psi$$

holds for any $\boldsymbol{\alpha}$, and for $\boldsymbol{\alpha} = (1-\theta)\boldsymbol{\gamma}+\theta\boldsymbol{\beta}$ with $\theta \notin (0,1)$,

$$\mathbf{C}_{\boldsymbol{\Delta}_\psi(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{C}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma})$$

follows from the restricted linearity of q. In particular, when $\nabla\psi(\boldsymbol{\alpha})$ exists, q is linear and so $\mathbf{C}_{q(\cdot;\boldsymbol{\alpha})} \equiv 0$ which gives the result in (ii).

A.6. Proof of Theorem 1. The theorem can be proved based on Theorem 6.1 of [37] and property (iii) of Lemma 3 in Section 2.1. We give some details for the second conclusion; the proof of the first follows similar lines and is easier. Consider a signal subclass

$$\mathcal{B}^1 = \{\boldsymbol{\beta} : \beta_j \in \{0,\tau R\}, \|\boldsymbol{\beta}\|_0 \leq s^*\},$$

where

$$R = [\sigma(\log(ep/s^*))^{1/2}/\overline{\kappa}^{1/2}] \wedge M$$

and $1 > \tau > 0$ is a small constant to be chosen later. Clearly, $\mathcal{B}^1 \in \mathcal{B}(s^*, M)$. By Stirling's approximation, $\log|\mathcal{B}^1| \geq \log\binom{p}{s^*} \geq s^*\log(p/s^*) \geq cs^*\log(ep/s^*)$ for some universal constant c.

Let $\rho(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) = \|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_0$, the Hamming distance between $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. By Lemma A.3 in [44], there exists a subset $\mathcal{B}^{10} \subset \mathcal{B}^1$ such that $\mathbf{0} \in \mathcal{B}^{10}$ and

$$\log|\mathcal{B}^{10}| \geq c_1 s^*\log(ep/s^*), \rho(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \geq c_2 s^*, \forall \boldsymbol{\beta}_1,\boldsymbol{\beta}_2 \in \mathcal{B}^{10}, \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$$

for some universal constants $c_1, c_2 > 0$. Then

$$\|\boldsymbol{X}\boldsymbol{\beta}_1 - \boldsymbol{X}\boldsymbol{\beta}_2\|_2^2 \geq \underline{\kappa}\|\boldsymbol{\beta}_1-\boldsymbol{\beta}_2\|_2^2 = \underline{\kappa}\tau^2R^2\rho(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2) \geq c_2\underline{\kappa}\tau^2R^2s^* \tag{A.6}$$

for any $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2 \in \mathcal{B}^{10}$, $\boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2$.

By Lemma 3 (iii), since Ω is open, for any $\boldsymbol{\beta} \in \mathcal{B}^{10}$, we have

$$\mathrm{KL}(p_{\boldsymbol{\beta}}, p_{\mathbf{0}}) = \boldsymbol{\Delta}_{l_0}(\mathbf{0}, \boldsymbol{X}\boldsymbol{\beta}) \le \tau^2 \overline{\kappa} R^2 s^* / (2\sigma^2).$$

Therefore,

$$\frac{1}{|\mathcal{B}^{10}| - 1} \sum_{\boldsymbol{\beta} \in \mathcal{B}^{10} \setminus \{\mathbf{0}\}} \mathrm{KL}(p_{\boldsymbol{\beta}}, p_{\mathbf{0}}) \le \tau^2 s^* \log(ep/s^*). \tag{A.7}$$

Combining (A.6) and (A.7) and choosing a sufficiently small value for τ, we can apply Theorem 2.7 of [58] to get the desired lower bound.

A.7. Proof of Proposition 1. We first introduce a lemma.

LEMMA A.2. *For the sequence of iterates $\{\boldsymbol{\beta}^{(t)}\}$ defined by (31) starting from an arbitrary point $\boldsymbol{\beta}^{(0)}$, if $f(\cdot)$ and $g(\cdot; \boldsymbol{\beta}^{(t)})$ are directionally differentiable, the following inequality holds for any $\boldsymbol{\beta}$ and $t \ge 0$*

$$\begin{aligned} &f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) \\ &\ge f(\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + (\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot; \boldsymbol{\beta}^{(t)})} + \boldsymbol{\Delta}_f)(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}). \end{aligned} \tag{A.8}$$

It can be proved by Lemma A.1 and Lemma 1 (details omitted). Rearranging (A.8) gives

$$\begin{aligned} &f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}) \\ &\le \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) - \boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot; \boldsymbol{\beta}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}). \end{aligned}$$

Under $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}) \ge 0$, we have

$$f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) \le \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) - \boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot; \boldsymbol{\beta}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}). \tag{A.9}$$

By Lemma 2, when ψ is differentiable, $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot; \boldsymbol{\beta}^{(t)})}$ is well-defined and equals $\boldsymbol{\Delta}_{\psi}$. Adding up the corresponding inequality for $t = 0, 1, \ldots, T$ leads to

$$\sum_{t=0}^{T} [f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta})] \le \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) - \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(T+1)}).$$

Therefore,

$$\operatorname*{avg}_{0 \le t \le T} f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) \le \frac{1}{T+1} [\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) - \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(T+1)})].$$

Note that under just the directional differentiability of $\boldsymbol{\Delta}_{\psi}(\cdot; \boldsymbol{\beta}^{(t)})$, (35) can be replaced by $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + (\boldsymbol{\Delta}_f + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot; \boldsymbol{\beta}^{(t)})} - \boldsymbol{\Delta}_{\psi})(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}) \ge 0,\ 0 \le t \le T$.

In the specific case that both f and ψ are convex, (35) is always satisfied by Lemma 1 and letting $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ in (A.9) gives

$$f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}^{(t)}) \le -\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le 0.$$

Hence $f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) = \min_{0 \le t \le T} f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) \le \operatorname{avg}_{0 \le t \le T} f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta})$. The proof is complete.

A.8. Proof of Proposition 2. Substituting $\boldsymbol{\beta}^o$ for $\boldsymbol{\beta}$ in Lemma A.2 gives

$$\begin{aligned} &f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}^o) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})}(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}) \\ &\le \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t)}) - \boldsymbol{\Delta}_f(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}). \end{aligned} \tag{A.10}$$

By Lemma A.1, we get

$$f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}^o) \ge \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^o). \tag{A.11}$$

Combining (A.10) and (A.11) yields

$$(2\bar{\boldsymbol{\Delta}}_f + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})})(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \le \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t)}). \tag{A.12}$$

It follows from the strong idempotence property that

$$(2\bar{\boldsymbol{\Delta}}_f + \boldsymbol{\Delta}_\psi)(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}) \le \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t)}) - \min_{0\le t\le T} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}), \tag{A.13}$$

for any $0 \le t \le T$, and so (40) can be obtained under $2\bar{\boldsymbol{\Delta}}_f \ge \varepsilon \boldsymbol{\Delta}_\psi$.

To show the first result, since $\boldsymbol{\Delta}_\phi = \boldsymbol{\Delta}_\psi + \boldsymbol{\Delta}_f$, (A.13) becomes

$$(2\bar{\boldsymbol{\Delta}}_f + \boldsymbol{\Delta}_\phi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}) \le (\boldsymbol{\Delta}_\phi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t)}) - \min_{0\le t\le T} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}).$$

Because $\kappa > 1$, (37) implies that

$$\boldsymbol{\Delta}_f \ge (\kappa+1)\bar{\boldsymbol{\Delta}}_\phi/\kappa - \grave{\boldsymbol{\Delta}}_\phi.$$

Applying the inequality twice, we obtain $((\kappa+1)/\kappa)\bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}) \le (2-(\kappa+1)/\kappa)\bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t)}) - \min_{0\le t\le T} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)})$, or

$$\bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t+1)}) \le \frac{\kappa-1}{\kappa+1}\bar{\boldsymbol{\Delta}}_\phi(\boldsymbol{\beta}^o, \boldsymbol{\beta}^{(t)}) - \frac{\kappa}{\kappa+1}\min_{0\le t\le T} \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}). \tag{A.14}$$

The final conclusion can be obtained by applying (A.14) iteratively for $t = 0, 1, \ldots, T$.

A.9. Proofs of Theorem 2 and Corollary 1. The proof of the theorem follows from Section A.7. In fact, setting $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t)}$ in (A.8) gives

$$(\grave{\boldsymbol{\Delta}}_\psi + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})} + \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}^{(t+1)}),$$

which, by the weak idempotence property (with $\boldsymbol{\alpha} = \boldsymbol{\beta}^{(t)}$), reduces to

$$(2\bar{\boldsymbol{\Delta}}_\psi + \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}^{(t+1)}). \tag{A.15}$$

Summing up (A.15) over $t = 0, 1, \ldots, T$ gives the conclusion.

Next, we prove a result slightly more general than Corollary 1. Recall the surrogate

$$g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + (\rho \mathbf{D}_\varphi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$$

where $\varphi \in \mathcal{C}^1$ is a strictly convex function, and f is continuous and directionally differentiable but not necessarily convex or differentiable. Denote $\arg\min g(\boldsymbol{\beta}; \boldsymbol{\beta}^-)$ by $\mathcal{T}(\boldsymbol{\beta}^-)$.

COROLLARY 1'. *Suppose that $\boldsymbol{\Delta}_f \le L\bar{\mathbf{D}}_\varphi$ for some $L > 0$ and the inverse stepsize parameter ρ satisfies $\rho > L/2$. Then $\operatorname{avg}_{0\le t\le T}(2\rho\bar{\mathbf{D}}_\varphi - \grave{\boldsymbol{\Delta}}_f)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le \frac{f(\boldsymbol{\beta}^{(0)})}{(T+1)}$ and so $\operatorname{avg}_{0\le t\le T} \bar{\mathbf{D}}_\varphi(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \le \frac{f(\boldsymbol{\beta}^{(0)})}{(T+1)(2\rho-L)}$.*

Moreover, for any accumulation point of $\boldsymbol{\beta}^{(t)}$ at which $\mathcal{T}$ is continuous, it must be a fixed point of $\mathcal{T}$. This is particularly true when $f \in \mathcal{C}^1$.

PROOF. Observe from (A.15) that

$$\begin{aligned} f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}^{(t+1)}) &\geq (2\rho\bar{\mathbf{D}}_\varphi - 2\bar{\boldsymbol{\Delta}}_f + \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \\ &\geq (2\rho\bar{\mathbf{D}}_\varphi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \\ &\geq (2\rho - L)\bar{\mathbf{D}}_\varphi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \geq 0. \end{aligned}$$

The error bounds can be obtained.

Let $\boldsymbol{\beta}^o$ be the limit point of some subsequence $\boldsymbol{\beta}^{t_l}$ as $l \to \infty$. Hence $f(\boldsymbol{\beta}^{(t)})$ converges monotonically to $\lim_{l\to\infty} f(\boldsymbol{\beta}^{t_l}) = f(\boldsymbol{\beta}^o)$. It follows that

$$\lim_{t\to+\infty} \bar{\mathbf{D}}_\varphi(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\beta}^{(t)}) = 0.$$

$\mathcal{T}$ is a well-defined function because of the strict convexity of the g-optimization problem. From the continuity assumptions,

$$0 = \lim_{l\to+\infty} \bar{\mathbf{D}}_\varphi(\boldsymbol{\beta}^{(t_l+1)}, \boldsymbol{\beta}^{(t_l)}) = \bar{\mathbf{D}}_\varphi(\mathcal{T}(\boldsymbol{\beta}^o), \boldsymbol{\beta}^o)$$

and thus $\mathcal{T}(\boldsymbol{\beta}^o) = \boldsymbol{\beta}^o$, i.e., $\boldsymbol{\beta}^o$ is a fixed point of $\mathcal{T}$. □

A.10. Proof of Proposition 3. First, we show a result when using the Bregman surrogate $g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$ for solving $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta})$ where l and P directionally differentiable and can be nonconvex. Define

$$\mathcal{L}_P := \inf\{\mathcal{L} \in \mathbb{R} : \boldsymbol{\Delta}_P + \mathcal{L}\mathbf{D}_2 \geq 0\}, \tag{A.16}$$

which provides an index to characterize the degree of nonconvexity of P, c.f. [36]. Assume $\mathcal{L}_P > -\infty$. Then for $\boldsymbol{\beta}^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)})$, the following inequality holds for all $T \geq 1$

$$\operatorname*{avg}_{0\leq t\leq T} (2\bar{\boldsymbol{\Delta}}_\psi + \boldsymbol{\Delta}_l - \varrho^2\mathcal{L}_P\mathbf{D}_2)(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \leq \frac{1}{T+1}\big[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta}^{(T+1)})\big].$$

The result can be proved from Theorem 2, noticing the fact that $\boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^-) = \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^-) + \boldsymbol{\Delta}_P(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^-) \geq \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^-) - \mathcal{L}_P\mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^-) = \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^-) - \varrho^2\mathcal{L}_P\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$ for any $\boldsymbol{\beta}, \boldsymbol{\beta}^-$. The details are omitted.

It suffices to proving the following lemma to complete the proof of Proposition 3.

LEMMA A.3. *Given any thresholding function Θ satisfying Definition 3, let P_Θ be the Θ-induced penalty in (48). Then $\mathcal{L}_\Theta$ as defined in (47) equals $\mathcal{L}_{P_\Theta}$ that is given in (A.16).*

PROOF. Since $\boldsymbol{\Delta}_{P_\Theta}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_j \boldsymbol{\Delta}_{P_\Theta}(\beta_j, \gamma_j)$, it suffices to show the result in the univariate case. Recall that $\Theta^{-1}(u;\lambda) := \sup\{t : \Theta(t;\lambda) \leq u\}, \forall u > 0$. Since $P_\Theta(\gamma) = P_\Theta(|\gamma|) = \int_0^{|\gamma|}(\Theta^{-1}(u;\lambda) - u)\,\mathrm{d}u$, we assume $\gamma \geq 0$ without loss of generality. We define $s(u;\lambda) = \Theta^{-1}(u;\lambda) - u$ for $u \geq 0$, and extend $s(\cdot)$ to $(-\infty, 0)$ by $s(-u) = -s(u), u > 0$. Clearly, $s'(u) = s'(|u|)$ a.e., and so $-\mathcal{L}_\Theta = \operatorname{ess\,inf}\{s'(u;\lambda) : u \neq 0\}$. By definition,

$$\delta P_\Theta(\gamma; \beta - \gamma) = \begin{cases} s(\gamma)(\beta - \gamma), & \text{if } \gamma \geq 0, \\ s(0)|\beta|, & \text{if } \gamma = 0. \end{cases}$$

When $\beta \geq 0$ and $\gamma \neq 0$, we get

$$\begin{aligned}
&(\boldsymbol{\Delta}_{P_\Theta} + \mathcal{L}\mathbf{D}_2)(\beta, \gamma) \\
&= P_\Theta(\beta) - P_\Theta(\gamma) - \delta P_\Theta(\gamma; \beta - \gamma) + \mathcal{L}\mathbf{D}_2(\beta, \gamma) \\
&= \int_\gamma^\beta s(u)\,\mathrm{d}u - s(\gamma)(\beta - \gamma) + \frac{1}{2}\mathcal{L}(\beta - \gamma)^2 \\
&= \int_\gamma^\beta s(u)\,\mathrm{d}u - \int_\gamma^\beta s(\gamma)\,\mathrm{d}u + \mathcal{L}\int_\gamma^\beta (u - \gamma)\,\mathrm{d}u \\
&= \int_\gamma^\beta \big[s(u) - s(\gamma) + \mathcal{L}(u - \gamma)\big]\,\mathrm{d}u \\
&= \int_\gamma^\beta \int_\gamma^u \big[s'(v) + \mathcal{L}\big]\,\mathrm{d}v\,\mathrm{d}u.
\end{aligned}$$

When $\beta < 0$ and $\gamma \neq 0$,

$$\begin{aligned}
&(\boldsymbol{\Delta}_{P_\Theta} + \mathcal{L}\mathbf{D}_2)(\beta, \gamma) \\
&= P_\Theta(\beta) - P_\Theta(\gamma) - \delta P_\Theta(\gamma; \beta - \gamma) + \mathcal{L}\mathbf{D}_2(\beta, \gamma) \\
&= \int_\gamma^{-\beta} s(u)\,\mathrm{d}u - s(\gamma)(\beta - \gamma) + \frac{1}{2}\mathcal{L}(\beta - \gamma)^2 \\
&= \int_{-\gamma}^{-\beta} s(u)\,\mathrm{d}u - \int_{-\gamma}^{-\beta} s(-\gamma)\,\mathrm{d}u + \mathcal{L}\int_{-\gamma}^{-\beta} (u + \gamma)\,\mathrm{d}u \\
&= \int_{-\gamma}^{-\beta} \big[s(u) - s(-\gamma) + \mathcal{L}(u + \gamma)\big]\,\mathrm{d}u \\
&= \int_{-\gamma}^{-\beta} \int_{-\gamma}^{u} \big[s'(v) + \mathcal{L}\big]\,\mathrm{d}v\,\mathrm{d}u.
\end{aligned}$$

Similarly, when $\gamma = 0$, $(\boldsymbol{\Delta}_{P_\Theta} + \mathcal{L}\mathbf{D}_2)(\beta, 0) = \int_0^{|\beta|} \int_0^u [s(v) + \mathcal{L}]\,\mathrm{d}v\,\mathrm{d}u$. It is then easy to verify that $\mathcal{L}_\Theta = \mathcal{L}_{P_\Theta}$. □

A.11. Proof of Proposition 4. Let $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta})$ and recall

$$\begin{aligned}
g^{(t)}_{\mathrm{LLA}}(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) &= f(\boldsymbol{\beta}) + \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^{(t)}) \\
&= f(\boldsymbol{\beta}) + \sum_j (\alpha_j^{(t)} \boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_P)(\varrho\beta_j, \varrho\beta_j^{(t)}).
\end{aligned}$$

The proof is similar to that of Theorem 2 and we give some details for completeness. The important fact $\beta^{(t+1)} \in \arg\min_{\boldsymbol{\beta}} g^{(t)}_{\mathrm{LLA}}(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)})$ as shown in Example 5 implies $\boldsymbol{\Delta}_{g^{(t)}_{\mathrm{LLA}}(\cdot; \boldsymbol{\beta}^{(t)})}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) \leq f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t)}, \varrho\boldsymbol{\beta}^{(t)}) - \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t+1)}, \varrho\boldsymbol{\beta}^{(t)})$

or

$$
\begin{aligned}
&f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}^{(t+1)}) \\
\geq\ & \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_{\boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\cdot, \varrho\boldsymbol{\beta}^{(t)})}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t+1)}, \varrho\boldsymbol{\beta}^{(t)}) \\
=\ & \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_{\boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\cdot, \varrho\boldsymbol{\beta}^{(t)})}(\varrho\boldsymbol{\beta}^{(t)}, \varrho\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t+1)}, \varrho\boldsymbol{\beta}^{(t)}) \\
=\ & \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t)}, \varrho\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t+1)}, \varrho\boldsymbol{\beta}^{(t)}). \\
=\ & \boldsymbol{\Delta}_f(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) + 2\bar{\boldsymbol{\Delta}}^{(t)}_{\mathrm{LLA}}(\varrho\boldsymbol{\beta}^{(t)}, \varrho\boldsymbol{\beta}^{(t+1)}).
\end{aligned}
$$

The conclusion follows from summing up this inequality for $t = 0, 1, \ldots, T$.

A.12. Proofs of Theorem 3 and Theorem 4. Let $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda)$ and recall $g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \boldsymbol{\beta}^-)$. We first introduce a lemma.

LEMMA A.4. *Let $\hat{\boldsymbol{\beta}} \in \mathcal{F}$. Then for any $\boldsymbol{\beta} \in \mathbb{R}^p$, we have the following inequality regardless of the specific form of ψ*

$$
\begin{aligned}
&(\boldsymbol{\Delta}_l - \varrho^2\mathcal{L}_\Theta\mathbf{D}_2)(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + P_\Theta(\varrho\hat{\boldsymbol{\beta}}; \lambda) \\
\leq\ & \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \langle \boldsymbol{\epsilon}, \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\rangle + P_\Theta(\varrho\boldsymbol{\beta}; \lambda).
\end{aligned} \tag{A.17}
$$

PROOF. Denote $\hat{g}(\boldsymbol{\beta}) := g(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}) = l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$. Since $\hat{\boldsymbol{\beta}}$ is a minimizer of $\hat{g}(\cdot)$, Lemma A.1 shows that for any $\boldsymbol{\beta}$, $\boldsymbol{\Delta}_{\hat{g}}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \leq \hat{g}(\boldsymbol{\beta}) - \hat{g}(\hat{\boldsymbol{\beta}})$. On the one hand,

$$
\begin{aligned}
&\hat{g}(\boldsymbol{\beta}) - \hat{g}(\hat{\boldsymbol{\beta}}) \\
=\ & l(\boldsymbol{\beta}) - l(\hat{\boldsymbol{\beta}}) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) - P_\Theta(\varrho\hat{\boldsymbol{\beta}}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \\
=\ & l(\boldsymbol{\beta}) - l(\boldsymbol{\beta}^*) - (l(\hat{\boldsymbol{\beta}}) - l(\boldsymbol{\beta}^*)) + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) - P_\Theta(\varrho\hat{\boldsymbol{\beta}}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \\
=\ & \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \langle \nabla l(\boldsymbol{\beta}^*), \boldsymbol{\beta} - \boldsymbol{\beta}^*\rangle - (\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \langle \nabla l(\boldsymbol{\beta}^*), \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\rangle) \\
&+ P_\Theta(\varrho\boldsymbol{\beta}; \lambda) - P_\Theta(\varrho\hat{\boldsymbol{\beta}}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \\
=\ & \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*) - \boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) + \langle \boldsymbol{\epsilon}, \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\rangle + P_\Theta(\varrho\boldsymbol{\beta}; \lambda) - P_\Theta(\varrho\hat{\boldsymbol{\beta}}; \lambda) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}).
\end{aligned}
$$

On the other hand, by Lemma 1, Lemma 2, and Lemma A.3,

$$
\begin{aligned}
\boldsymbol{\Delta}_{\hat{g}}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) &= \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_{P_\Theta(\varrho\cdot)}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot, \hat{\boldsymbol{\beta}})}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \\
&= \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_{P_\Theta(\varrho\cdot)}(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \\
&\geq \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) - \varrho^2\mathcal{L}_\Theta\mathbf{D}_2(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}, \hat{\boldsymbol{\beta}}).
\end{aligned}
$$

The conclusion follows. □

To handle the stochastic term $\langle \boldsymbol{\epsilon}, \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\rangle$ in (A.17), we introduce the following result.

LEMMA A.5. *Let $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{\epsilon}$ be a sub-Gaussian random vector with mean 0 and scale bounded by σ, and $\lambda^o = \sigma\sqrt{\log(ep)}$. Suppose that $\varrho \geq \|\boldsymbol{X}\|_2$. Then there exist universal constants $A_0, C, c > 0$ such that for any $a \geq 2b > 0$ and $A_1 \geq A_0$, the following event*

$$
\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} \left\{ 2\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{a}\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{b}[P_H(\varrho\boldsymbol{\beta}; \sqrt{ab}A_1\lambda^o)] \right\} \geq a\sigma^2 t
$$

occurs with probability at most $C\exp(-ct)p^{-cA_1^2}$.

The lemma can be proved by Lemma 4 of [49] based on a scaling argument.

Let $R = \sup_{\boldsymbol{\beta},\hat{\boldsymbol{\beta}}\in\mathbb{R}^p}\{\langle \boldsymbol{\epsilon}, \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\rangle - \frac{1}{2a}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 - \frac{1}{2b}[P_H(\varrho(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}); \sqrt{ab}A\lambda^o)]\}$ with $\lambda^o = \sigma\sqrt{\log(ep)}$. Plugging the bound into (A.17) gives

$$\begin{aligned}&(\boldsymbol{\Delta}_l - \varrho^2\mathcal{L}_\Theta\mathbf{D}_2)(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*) + P_\Theta(\varrho\hat{\boldsymbol{\beta}};\lambda) - P_\Theta(\varrho\boldsymbol{\beta};\lambda)\\ &\le \boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^*) + \frac{1}{2a}\|\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \frac{1}{2b}[P_H(\varrho(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}); \sqrt{ab}A\lambda^o)] + R\end{aligned} \tag{A.18}$$

with $\mathbb{P}(2R \ge a\sigma^2 t) \le C\exp(-ct)p^{-cA^2}$ for any $a \ge 2b > 0$ and A large.

To prove Theorem 3, substitute $\boldsymbol{\beta}^*$ for $\boldsymbol{\beta}$ in (A.18) and combine it with the regularity condition (57), resulting in

$$\big(\delta - \frac{1}{a}\big)\mathbf{D}_2(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \boldsymbol{X}\boldsymbol{\beta}^*) + \big(\vartheta - \frac{1}{2b}\big)P_H(\varrho(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*;\lambda) \le K\lambda^2 J(\boldsymbol{\beta}^*) + R$$

where $\lambda = \sqrt{ab}A\lambda^o$, $a \ge 2b > 0$, and $A \ge A_0$ with A_0 given in Lemma A.5. Setting $a = 2/(\delta \wedge (2\vartheta)), b = 1/(2\vartheta)$ or $a = 2/((2\delta) \wedge \vartheta), b = 1/\vartheta$ bounds $\mathbf{D}_2(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \boldsymbol{X}\boldsymbol{\beta}^*)$ or $P_H(\varrho(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*);\lambda)$. Finally, by Lemma A.5, $\mathbb{P}(R \le 0) \ge 1 - Cp^{-cA^2}$.

For Theorem 4, we combine (A.18) and (60) with $\boldsymbol{\gamma} = \hat{\boldsymbol{\beta}}$:

$$\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*) + \big(\delta - \frac{1}{a}\big)\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\hat{\boldsymbol{\beta}}) \le \frac{\alpha}{L}r\boldsymbol{\Delta}_l(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) + \boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^*) + K\lambda^2 J(\boldsymbol{\beta}) + R.$$

Take the same choice for λ and set $a = 2/(\delta \wedge (2\vartheta)), b = 1/(2\vartheta)$.

Case (i): $\alpha r/L = 0$. The conclusion follows easily, and does not need any restriction on α or L.

Case (ii): $\alpha r/L > 0$. Then $0 < r < 1$ and $\alpha/L > 0$. If $\alpha < 0$, $(\alpha r/L)\boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\hat{\boldsymbol{\beta}}) \le \alpha r d^2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\hat{\boldsymbol{\beta}}) \le 0$, reducing to the first case. Assume $\alpha > 0$. Then

$$\begin{aligned}\frac{\alpha}{L}r\boldsymbol{\Delta}_l(\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) &= \frac{\alpha}{L}r\boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\hat{\boldsymbol{\beta}}) \le \alpha r d^2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\hat{\boldsymbol{\beta}})\\ &\le \alpha r(d(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}^*) + d(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \boldsymbol{X}\boldsymbol{\beta}^*))^2\\ &\le \alpha r(1 + 1/M)d^2(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}^*) + \alpha r(1 + M)d^2(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \boldsymbol{X}\boldsymbol{\beta}^*)\\ &\le r(1 + 1/M)\boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^*) + r(1 + M)\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}^*),\end{aligned}$$

for any $M > 0$. Take $M = (1 - r)/(1 + r)$. Then

$$1/\{1 - r(1 + M)\} = 1 + r(1 + 1/M) = (1 + r)/(1 - r).$$

So we obtain

$$\begin{aligned}&(\frac{1 - r}{1 + r}\boldsymbol{\Delta}_{l_0} + \frac{\delta}{2}\mathbf{D}_2)(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \boldsymbol{X}\boldsymbol{\beta}^*)\\ &\le \frac{1 + r}{1 - r}\boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}^*) + \frac{KA^2}{((2\vartheta) \wedge \delta)\vartheta}\sigma^2 J(\boldsymbol{\beta})\log(ep) + R.\end{aligned}$$

Finally, from $\mathbb{P}(2R \ge a\sigma^2 t) \le C\exp(-ct)p^{-cA^2} \le C\exp(-ct)$, we have $\mathbb{E}R \le Ca\sigma^2$. The oracle inequality is proved. In fact, we also get

$$\begin{aligned}\mathbb{E}[\mathbf{D}_2(\boldsymbol{X}\hat{\boldsymbol{\beta}}, \boldsymbol{X}\boldsymbol{\beta}^*)] &\le \frac{2(1 + r)}{(1 - r)\delta}\mathbb{E}[\boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{X}\boldsymbol{\beta}^*)]\\ &+ \frac{2KA^2}{(\vartheta \wedge \delta)\vartheta\delta}\sigma^2 J(\boldsymbol{\beta})\log(ep) + \frac{C}{(\vartheta \wedge \delta)\delta}\sigma^2\end{aligned}$$

under the same condition.

REMARK A.1. *Recall $J^* = J(\boldsymbol{\beta}^*), \mathcal{J}^* = \mathcal{J}(\boldsymbol{\beta}^*)$ and $P(\boldsymbol{\beta};\lambda) = \sum_j P(\beta_j;\lambda)$. When $\hat{\boldsymbol{\beta}}$ is a global minimizer, applying the bound of the stochastic term proved in Lemma A.5 gives the same conclusions (58), (59), under*

$$\begin{aligned}&\vartheta P_H(\varrho(\boldsymbol{\beta}-\boldsymbol{\beta}^*);\lambda) + P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda)\\ &\le (2\bar{\boldsymbol{\Delta}}_{l_0} - \delta\mathbf{D}_2)(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{X}\boldsymbol{\beta}^*) + P_\Theta(\varrho\boldsymbol{\beta};\lambda) + K\lambda^2 J^*, \forall\boldsymbol{\beta}\end{aligned} \tag{A.19}$$

for some $\delta > 0$, $\vartheta > 0$ and large enough $K \ge 0$. Assuming P_Θ is subadditive, we can follow the arguments in Remark 2 to show that (A.19) is implied by $(1+\vartheta)P_\Theta(\varrho(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\mathcal{J}^*};\lambda) \le (2\bar{\boldsymbol{\Delta}}_{l_0} - \delta\mathbf{D}_2)(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{X}\boldsymbol{\beta}^*) + K\lambda^2 J^* + (1-\vartheta)P_\Theta(\varrho(\boldsymbol{\beta}-\boldsymbol{\beta}^*)_{\mathcal{J}^{*c}};\lambda)$. Furthermore, when l_0 is μ-strongly convex as in regression, one can take $\delta = \mu$ and the regularity condition is implied by

$$(1+\bar{\vartheta})P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^*};\lambda) \le \bar{K}\lambda\sqrt{J^*}\|\boldsymbol{X}\boldsymbol{\gamma}\|_2 + P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^{*c}};\lambda), \forall\boldsymbol{\gamma} \tag{A.20}$$

for some $\bar{\vartheta}\,(=\frac{2\vartheta}{1-\vartheta}) > 0$ and $\bar{K}\big(=\frac{\sqrt{2(2\mu-\delta)K}}{1-\vartheta}\big) \ge 0$, or the constrained forms

$$[P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^*};\lambda)]^2 \le \tilde{K}\lambda^2 J^*\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2,\ \forall\boldsymbol{\gamma} : P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^{*c}};\lambda) \le (1+\bar{\vartheta})P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^*};\lambda) \tag{A.21}$$

$$\sum_{j\in\mathcal{J}^*} P_\Theta^2(\varrho\gamma_j;\lambda) \le \tilde{K}\lambda^2\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2,\ \forall\boldsymbol{\gamma} : P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^{*c}};\lambda) \le (1+\bar{\vartheta})P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^*};\lambda) \tag{A.22}$$

for some $\bar{\vartheta} > 0$ and $\tilde{K} \ge 0$. The conclusions and conditions can also be formulated in the oracle inequality setup of Theorem 4. (A.20), (A.21) and (A.22) extend the comparison condition (62), compatibility condition and RE condition to a more general penalty.

A.13. Proof of Theorem 5. We prove the result under a more relaxed assumption: l, g are merely directionally differentiable, and (64) is replaced by

$$\begin{aligned}&(\varrho^2\mathcal{L}_\Theta\mathbf{D}_2 + \varepsilon\boldsymbol{\Delta}_\psi)(\boldsymbol{\beta}^*,\boldsymbol{\beta}) + (\boldsymbol{\Delta}_\psi - \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\alpha})})(\boldsymbol{\beta}^*,\boldsymbol{\beta}) + \vartheta P_H(\varrho(\boldsymbol{\beta}-\boldsymbol{\beta}^*);\lambda) + P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda)\\ &\le (2\bar{\boldsymbol{\Delta}}_{l_0} - \delta\mathbf{D}_2)(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{X}\boldsymbol{\beta}^*) + P_\Theta(\varrho\boldsymbol{\beta};\lambda) + K\lambda^2 J(\boldsymbol{\beta}^*), \forall\boldsymbol{\beta},\boldsymbol{\alpha}\end{aligned}$$

for some $\delta > 0$, $\varepsilon > 0$, $\vartheta > 0$ and large $K \ge 0$.

Recall the objective function $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P_\Theta(\varrho\boldsymbol{\beta};\lambda)$ and the surrogate function $g(\boldsymbol{\beta};\boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\beta}^-)$. From Lemma A.2 and Lemma A.3, we obtain

$$\begin{aligned}&f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta},\boldsymbol{\beta}^{(t)})\\ &\ge f(\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})} + \boldsymbol{\Delta}_l + \boldsymbol{\Delta}_{P_\Theta(\varrho\cdot)})(\boldsymbol{\beta},\boldsymbol{\beta}^{(t+1)})\\ &\ge f(\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})} + \boldsymbol{\Delta}_l - \varrho^2\mathcal{L}_\Theta\mathbf{D}_2)(\boldsymbol{\beta},\boldsymbol{\beta}^{(t+1)}).\end{aligned}$$

Substituting $\boldsymbol{\beta}^*$ for $\boldsymbol{\beta}$ yields

$$\begin{aligned}&\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t)})\\ &\ge \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)}) + (\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})} + \boldsymbol{\Delta}_l)(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)}) - \mathcal{L}_\Theta\mathbf{D}_2(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t+1)})\\ &\quad + l(\boldsymbol{\beta}^{(t+1)}) - l(\boldsymbol{\beta}^*) + P_\Theta(\varrho\boldsymbol{\beta}^{(t+1)};\lambda) - P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda)\\ &= \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)}) + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})}(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)}) - \mathcal{L}_\Theta\mathbf{D}_2(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t+1)})\\ &\quad 2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)}) - \langle\boldsymbol{\epsilon},\boldsymbol{X}\boldsymbol{\beta}^{(t+1)} - \boldsymbol{X}\boldsymbol{\beta}^*\rangle + P_\Theta(\varrho\boldsymbol{\beta}^{(t+1)};\lambda) - P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda).\end{aligned}$$

From the above regularity condition,

$$\begin{aligned}&\varepsilon\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})+\mathcal{L}_\Theta\mathbf{D}_2(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t+1)})+\delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*,\boldsymbol{X}\boldsymbol{\beta}^{(t+1)})\\&+\vartheta P_H(\varrho(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*);\lambda)+P_\Theta(\varrho\boldsymbol{\beta}^*;\lambda)+(\boldsymbol{\Delta}_\psi-\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\psi(\cdot;\boldsymbol{\beta}^{(t)})})(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})\\\le&\,2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})+P_\Theta(\varrho\boldsymbol{\beta}^{(t+1)};\lambda)+K\lambda^2J(\boldsymbol{\beta}^*),\end{aligned}$$

and so we obtain

$$\begin{aligned}&(\varepsilon+1)\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})+\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)})\\&+\delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*,\boldsymbol{X}\boldsymbol{\beta}^{(t+1)})+\vartheta P_H(\varrho(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*);\lambda)\\\le&\,\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t)})+\langle\boldsymbol{\epsilon},\boldsymbol{X}\boldsymbol{\beta}^{(t+1)}-\boldsymbol{X}\boldsymbol{\beta}^*\rangle+K\lambda^2J^*.\end{aligned}\tag{A.23}$$

According to Lemma A.5, as long as $A\ge A_0/\sqrt{2}$, choosing $\lambda=\sqrt{2ab}A\lambda^o$, $b=1/(2\vartheta), a=1/(\delta\wedge\vartheta)$ guarantees that the probability of the following inequality occurring for all t,

$$\langle\boldsymbol{\epsilon},\boldsymbol{X}\boldsymbol{\beta}^{(t+1)}-\boldsymbol{X}\boldsymbol{\beta}^*\rangle-\delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*,\boldsymbol{X}\boldsymbol{\beta}^{(t+1)})-\vartheta P_H(\varrho(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*);\lambda)\ge0,$$

is no greater than Cp^{-cA^2}. Together with (A.23), with probability $1-Cp^{-cA^2}$

$$\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})\le\frac{1}{\varepsilon+1}[\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t)})-\boldsymbol{\Delta}(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)})+K\lambda^2J^*]\tag{A.24}$$

for all t. The desired inequality can be shown by iteratively applying (A.24) for $t=0,1,2,\ldots$.

A.14. Proofs of Theorem 6 and Corollary 2. First we prove Theorem 6. Note that (67b), (69b) have additional terms involving μ_0. The first result for $\mu_0=0$ can be shown based on a GBF translation of the proof of Proposition 1 in [57]. For convenience, let $h_t(\boldsymbol{\beta})=f(\boldsymbol{\beta})-\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)})=f(\boldsymbol{\beta})-\boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)})+\mu_0\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)})$. Applying Lemma A.2 to (67b) yields $(\boldsymbol{\Delta}_f-\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot,\boldsymbol{\gamma}^{(t)})}+\theta_t\rho_t\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})})(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})\le h_t(\boldsymbol{\beta})+\theta_t\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})-h_t(\boldsymbol{\alpha}^{(t+1)})-\theta_t\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\alpha}^{(t)})\ \forall\boldsymbol{\beta}$, or

$$\begin{aligned}&h_t(\boldsymbol{\alpha}^{(t+1)})-h_t(\boldsymbol{\beta})+\theta_t\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\alpha}^{(t)})\\\le&\,\theta_t\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})-(\theta_t\rho_t\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})}+\boldsymbol{\Delta}_{f(\cdot)-\boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot,\boldsymbol{\gamma}^{(t)})})(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)}).\end{aligned}\tag{A.25}$$

By definition, $\mathbf{C}_{h_t}(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\beta}^{(t)},\theta_t)=\theta_th_t(\boldsymbol{\alpha}^{(t+1)})+(1-\theta_t)h_t(\boldsymbol{\beta}^{(t)})-h_t(\boldsymbol{\beta}^{(t+1)})$; adding it to (A.25) multiplied by θ_t gives

$$\begin{aligned}&h_t(\boldsymbol{\beta}^{(t+1)})-(1-\theta_t)h_t(\boldsymbol{\beta}^{(t)})-\theta_th_t(\boldsymbol{\beta})\\&+\theta_t^2\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\alpha}^{(t)})+\mathbf{C}_{h_t}(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\beta}^{(t)},\theta_t)\\&+\theta_t^2\rho_t\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})-\phi(\cdot)}(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})\\\le&\,\theta_t^2\rho_t\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})-(\theta_t^2\rho_t\boldsymbol{\Delta}_\phi+\theta_t\boldsymbol{\Delta}_{f(\cdot)-\boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot,\boldsymbol{\gamma}^{(t)})})(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)}),\end{aligned}$$

and so

$$\begin{aligned}&f(\boldsymbol{\beta}^{(t+1)})-f(\boldsymbol{\beta})-(1-\theta_t)[f(\boldsymbol{\beta}^{(t)})-f(\boldsymbol{\beta})]+\theta_t\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)})\\&+\theta_t\{(\boldsymbol{\Delta}_{f(\cdot)-\boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot,\boldsymbol{\gamma}^{(t)})}+\theta_t\rho_t\boldsymbol{\Delta}_{\boldsymbol{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})-\phi(\cdot)})(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})\}+R_t\\\le&\,\theta_t^2\rho_t(\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})-\boldsymbol{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})),\forall t\ge0\end{aligned}\tag{A.26}$$

where R_t is given by $\theta_t^2\rho_t\mathbf{\Delta}_\phi(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\alpha}^{(t)})-\mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\gamma}^{(t)})+(1-\theta_t)\mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)},\boldsymbol{\gamma}^{(t)})+\mathbf{C}_{f(\cdot)-\mathbf{\Delta}_{\bar{\psi}_0}(\cdot,\boldsymbol{\gamma}^{(t)})}(\boldsymbol{\alpha}^{(t+1)},\boldsymbol{\beta}^{(t)},\theta_t)$.

Under $\mu_0=0$, (A.26) implies that

$$\begin{aligned}(\text{A.27})\qquad &\frac{1}{\theta_t^2\rho_t}\big[f(\boldsymbol{\beta}^{(t+1)})-f(\boldsymbol{\beta})\big]-\frac{1-\theta_t}{\theta_t^2\rho_t}\big[f(\boldsymbol{\beta}^{(t)})-f(\boldsymbol{\beta})\big]+\frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t\rho_t}+\frac{R_t}{\theta_t^2\rho_t}\\ &\le\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})-\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)}).\end{aligned}$$

Since in this case (69b) gives $(1-\theta_t)/\theta_t^2\rho_t=1/\theta_{t-1}^2\rho_{t-1}$ for any $t\ge1$, we obtain the first conclusion

$$\begin{aligned}&\frac{1}{\theta_T^2\rho_T}\big[f(\boldsymbol{\beta}^{(T+1)})-f(\boldsymbol{\beta})\big]-\frac{1-\theta_0}{\theta_0^2\rho_0}\big[f(\boldsymbol{\beta}^{(0)})-f(\boldsymbol{\beta})\big]+\sum_{t=0}^{T}\Big(\frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t\rho_t}+\frac{R_t}{\theta_t^2\rho_t}\Big)\\ &\le\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(0)})-\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(T+1)}).\end{aligned}$$

On the other hand, given $\mu_0\ge0$, (A.26) can be written as

$$\begin{aligned}(\text{A.28})\qquad &f(\boldsymbol{\beta}^{(t+1)})-f(\boldsymbol{\beta})-(1-\theta_t)[f(\boldsymbol{\beta}^{(t)})-f(\boldsymbol{\beta})]\\ &+R_t+\theta_t\mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta},\boldsymbol{\gamma}^{(t)})+\theta_t\mathbf{\Delta}_{f(\cdot)-\mathbf{\Delta}_{\psi_0}(\cdot,\boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})\\ &+\theta_t(\mu_0\mathbf{\Delta}_{\mathbf{\Delta}_\phi(\cdot,\boldsymbol{\gamma}^{(t)})-\phi(\cdot)}+\theta_t\rho_t\mathbf{\Delta}_{\mathbf{\Delta}_\phi(\cdot,\boldsymbol{\alpha}^{(t)})-\phi(\cdot)})(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})\\ &\le\theta_t^2\rho_t\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})-\theta_t^2(\rho_t+\frac{\mu_0}{\theta_t})\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)}).\end{aligned}$$

Therefore, we have

$$\begin{aligned}&f(\boldsymbol{\beta}^{(t+1)})-f(\boldsymbol{\beta})+\theta_t^2(\rho_t+\frac{\mu_0}{\theta_t})\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})+\theta_t\mathcal{E}_t(\boldsymbol{\beta})+R_t\\ &\le(1-\theta_t)[f(\boldsymbol{\beta}^{(t)})-f(\boldsymbol{\beta})]+\theta_t^2\rho_t\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)}),\forall t\ge0\end{aligned}$$

and from (69b),

$$\begin{aligned}&f(\boldsymbol{\beta}^{(t+1)})-f(\boldsymbol{\beta})+\theta_t^2(\rho_t+\frac{\mu_0}{\theta_t})\mathbf{\Delta}_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t+1)})+\theta_t\mathcal{E}_t(\boldsymbol{\beta})+R_t\\ &\le(1-\theta_t)[f(\boldsymbol{\beta}^{(t)})-f(\boldsymbol{\beta})+\theta_{t-1}^2(\rho_{t-1}+\frac{\mu_0}{\theta_{t-1}})\Delta_\phi(\boldsymbol{\beta},\boldsymbol{\alpha}^{(t)})],\forall t\ge1\end{aligned}$$

The second conclusion can be obtained by a recursive argument and $R_T+\theta_T\mathcal{E}_T(\boldsymbol{\beta})+(1-\theta_T)(R_{T-1}+\theta_{T-1}\mathcal{E}_{T-1}(\boldsymbol{\beta}))+\cdots(1-\theta_T)\cdots(1-\theta_1)(R_0+\theta_0\mathcal{E}_0(\boldsymbol{\beta}))=\sum_{t=0}^{T}(\prod_{s=t+1}^{T}(1-\theta_s))(R_t+\theta_t\mathcal{E}_t(\boldsymbol{\beta}))$.

REMARK A.2. *With the '=' in (69b) replaced by '≤', (71) still holds when $\mathbf{\Delta}_\phi\ge0$ (or ϕ is convex), and (70) still holds if we set $\boldsymbol{\beta}$ to be a minimizer of f. But the equality form of (69b) makes our conclusions applicable to say the noise-free statistical truth $\boldsymbol{\beta}=\boldsymbol{\beta}^*$, which may not be a minimizer of the sample-based objective. The same comment applies to Theorem 7.*

Also, it is trivial to see that the conclusions extend to a varying sequence of μ_t. (Concretely, the μ_0 in (67b), (68), and $\mathcal{E}_t(\boldsymbol{\beta})$ becomes μ_t, and the μ_0 in (69b), (71) becomes μ_{t-1},μ_T, respectively.) One can add backtracking for μ_t in the algorithm to further reduce its iteration complexity.

Finally, we prove Corollary 2' which implies Corollary 2 and applies to any convex l_0 in (A.29) below. The proof is based on an accumulative R_t bound that can be derived in a more general setup; see (A.32) in Remark A.3.

Here, the optimization problem of interest in "variable screening" is

$$\min l(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta}) \text{ s.t. } \|\boldsymbol{\beta}\|_0 \le q, \tag{A.29}$$

to estimate the target $\boldsymbol{\beta}^*$ satisfying the strict inequality $\|\boldsymbol{\beta}^*\|_0 < q$. Take $\mu_0 = 0$, $\phi = \|\cdot\|_2^2/2$, $\psi_0(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) - \mathcal{L}\phi(\boldsymbol{\beta})$ for some $\mathcal{L} \ge 0$.

Given l_0, $\boldsymbol{X}$, and $s \le p$, we extend the notion of restricted isometry numbers ρ_+, ρ_- [12]:

$$\rho_-(s)\mathbf{D}_2(\boldsymbol{\beta},\boldsymbol{\gamma}) \le \boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{X}\boldsymbol{\gamma}) \le \rho_+(s)\mathbf{D}_2(\boldsymbol{\beta},\boldsymbol{\gamma}), \forall \boldsymbol{\beta},\boldsymbol{\gamma} : \|\boldsymbol{\beta}-\boldsymbol{\gamma}\|_0 \le s. \tag{A.30}$$

(The dependence on $\boldsymbol{X}$ and l_0 is dropped for the sake of brevity.)

COROLLARY 2'. *Let l_0 be any convex function and q be any nonnegative integer no more than p. As long as $r = q/\|\boldsymbol{\beta}^*\|_0 > 1$, for any $\mathcal{L} \ge \rho_+(2q)/\sqrt{r}$, there must exist a universal $\rho_t = \rho_0, \forall t$, e.g.,*

$$\rho_0 = (1 - 1/\sqrt{r})\rho_+(2q),$$

and thus $\theta_{t+1} = (\sqrt{\theta_t^4 + 4\theta_t^2} - \theta_t^2)/2$, so that the accelerated iterative quantile-thresholding according to (67a)–(67c) satisfies

$$\frac{l(\boldsymbol{\beta}^{(T+1)}) - l(\boldsymbol{\beta}^*)}{\theta_T^2} + T \cdot \operatorname*{avg}_{0\le t\le T} \frac{\boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)})}{\theta_t} + \frac{\mathcal{L}(1-\theta_T)}{\theta_T^3}\mathbf{D}_2(\boldsymbol{\beta}^{(T+1)}, \boldsymbol{\beta}^{(T)})$$
$$\le \rho_0 \mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(0)}), \text{ for all } T \ge 0,$$

or $l(\boldsymbol{\beta}^{(T+1)}) - l(\boldsymbol{\beta}^*) + \min_{0\le t\le T} \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)}) \le A/T^2$ with A independent of T.

Seen from the error measure, r should be appropriately large (but cannot be too large from the perspective of statistical accuracy.)

To prove the corollary, we first introduce a useful result [50, Lemma 9].

LEMMA A.6. *Given $q \le p$, $\hat{\boldsymbol{\beta}} = \Theta^{\#}(\boldsymbol{y}; q)$ is a globally optimal solution to $\min_{\boldsymbol{\beta}\in\mathbb{R}^p} l(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{\beta}\|_2^2/2$ s.t. $\|\boldsymbol{\beta}\|_0 \le q$. Let $\mathcal{J} = \mathcal{J}(\boldsymbol{\beta})$, $\hat{\mathcal{J}} = \mathcal{J}(\hat{\boldsymbol{\beta}})$ and assume $J(\hat{\boldsymbol{\beta}}) = q$. Then, for any $\boldsymbol{\beta}$ with $J(\boldsymbol{\beta}) \le s = q/r$ and $r \ge 1$,*

$$l(\boldsymbol{\beta}) - l(\hat{\boldsymbol{\beta}}) \ge \{1 - \mathcal{L}(\mathcal{J}, \hat{\mathcal{J}})\}\mathbf{D}_2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}),$$

where $\mathcal{L}(\mathcal{J}, \hat{\mathcal{J}}) = (|\mathcal{J}\backslash\hat{\mathcal{J}}|/|\hat{\mathcal{J}}\backslash\mathcal{J}|)^{1/2} \le (s/q)^{1/2} = r^{-1/2}$.

Set $\boldsymbol{\beta} = \boldsymbol{\beta}^*$ in the previous proof, and apply, instead of Lemma A.2, Lemma A.6 (where $\|\boldsymbol{\alpha}^{(t+1)}\|_0 = q$ due to the no-tie-occurring assumption) to (67b). (A.25) is then replaced by

$$l(\boldsymbol{\alpha}^{(t+1)}) - \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) - l(\boldsymbol{\beta}^*) + \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)}) + \theta_t\rho_t\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)})$$
$$\le \theta_t\rho_t\mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t)}) - (\theta_t\rho_t + \mathcal{L})(1 - \frac{1}{\sqrt{r}})\mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t+1)}).$$

Accordingly, (A.26) becomes

$$l(\boldsymbol{\beta}^{(t+1)}) - l(\boldsymbol{\beta}^*) - (1-\theta_t)(l(\boldsymbol{\beta}^{(t)}) - l(\boldsymbol{\beta}^*)) + R_t$$
$$+ \theta_t\Big\{\boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)}) + \big[\mathcal{L}(1 - \frac{1}{\sqrt{r}}) - \frac{\theta_t\rho_t}{\sqrt{r}}\big]\mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t+1)})\Big\}$$
$$\le \theta_t^2\rho_t(\mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t)}) - \mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t+1)})),$$

where R_t is the same as before. Based on Lemma 1, Lemma 2, Lemma 5 (together with some results in its proof), (A.30), and the following facts

$$\boldsymbol{\beta}^{(t)} - \boldsymbol{\gamma}^{(t)} = \theta_t(\boldsymbol{\beta}^{(t)} - \boldsymbol{\alpha}^{(t)}) = \theta_t(1-\theta_{t-1})(\boldsymbol{\beta}^{(t-1)} - \boldsymbol{\alpha}^{(t)}), \forall t \geq 1$$
$$\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\gamma}^{(t)} = \theta_t(\boldsymbol{\alpha}^{(t+1)} - \boldsymbol{\alpha}^{(t)}), \forall t \geq 0$$

we obtain for all $t \geq 1$,

$$\begin{aligned}
R_t \geq & \theta_t^2\rho_t \mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) - \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \\
& + \mathcal{L}\theta_t(1-\theta_t)\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \\
= & \theta_t^2(\rho_t + \mathcal{L})\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) - \boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}^{(t+1)}, \boldsymbol{X}\boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\boldsymbol{\Delta}_l(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \\
& + \mathcal{L}\{\theta_t(1-\theta_t)\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}) - (1-\theta_t)\mathbf{D}_2(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})\} \\
= & \theta_t^2(\rho_t + \mathcal{L})\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) - \boldsymbol{\Delta}_{l_0}(\boldsymbol{X}\boldsymbol{\beta}^{(t+1)}, \boldsymbol{X}\boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\boldsymbol{\Delta}_l(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \\
& + \mathcal{L}\{\theta_t(1-\theta_t)\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}) - (1-\theta_t)\theta_t^2(1-\theta_{t-1})^2\mathbf{D}_2(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t-1)})\} \\
\geq & \theta_t^2(\rho_t + \mathcal{L})\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) - \theta_t^2\rho_+(2q)\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) \\
& + \mathcal{L}\theta_t(1-\theta_t)\{\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}) - \theta_t(1-\theta_{t-1})^2\mathbf{D}_2(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t-1)})\},
\end{aligned}$$

and $R_t \geq \theta_t^2\{\rho_t + \mathcal{L} - \rho_+(2q)\}\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)})$ as $t = 0$. It follows that

$$\begin{aligned}
\sum_{t=0}^{T}\frac{R_t}{\theta_t^2\rho_t} \geq & \sum_{t=0}^{T}\frac{\rho_t + \mathcal{L} - \rho_+(2q)}{\rho_t}\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) + \mathcal{L}\frac{1-\theta_T}{\theta_T\rho_T}\mathbf{D}_2(\boldsymbol{\alpha}^{(T+1)}, \boldsymbol{\beta}^{(T)}) \\
& + \mathcal{L}\sum_{t=0}^{T-1}\Big\{\frac{1-\theta_t}{\theta_t\rho_t} - \frac{(1-\theta_t)^2(1-\theta_{t+1})}{\rho_{t+1}}\Big\}\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}) \\
= & \sum_{t=0}^{T}\frac{\rho_t + \mathcal{L} - \rho_+(2q)}{\rho_t}\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) + \frac{\mathcal{L}(1-\theta_T)}{\theta_T^3\rho_T}\mathbf{D}_2(\boldsymbol{\beta}^{(T+1)}, \boldsymbol{\beta}^{(T)}) \\
& + \mathcal{L}\sum_{t=0}^{T-1}\frac{1-\theta_t}{\theta_t\rho_t}\Big\{1 - (1-\theta_t)\theta_{t+1}\frac{\theta_{t+1}}{\theta_t}\Big\}\mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}).
\end{aligned}$$

Therefore, choosing a universal $\rho_t = \rho_0 \geq \rho_+(2q) - \mathcal{L}$ (which implies $\theta_t \downarrow$) ensures $\sum_{t=0}^{T} R_t/(\theta_t^2\rho_t) \geq 0$.

Moreover, $\mathcal{E}_t(\boldsymbol{\beta}^*) \geq \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)})$ holds under $\{\mathcal{L}(1-\frac{1}{\sqrt{r}}) - \frac{\theta_t\rho_t}{\sqrt{r}}\}\mathbf{D}_2(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^{(t+1)})\} \geq 0$ or $L(1-1/\sqrt{r}) \geq \rho_t/\sqrt{r}$. It is easy to see that as long as $r > 1$, there exist positive $\mathcal{L}, \rho_0$ satisfying

$$\begin{cases}\mathcal{L}(\sqrt{r}-1) \geq \rho_0 \\ \rho_0 \geq \rho_+(2q) - \mathcal{L}.\end{cases} \tag{A.31}$$

Furthermore, for any $\mathcal{L} \geq \rho_+(2q)/\sqrt{r}$, we can always choose $\rho_0 = (1-1/\sqrt{r})\rho_+(2q)$. The rest of the proof proceeds as before.

REMARK A.3. *The idea of controlling the overall* $\sum_{t\leq T}\frac{R_t}{\theta_t^2\rho_t}$ *can be extended with a proper choice of* ψ_0 *to a general problem* $\min f(\boldsymbol{\beta})$ *that may be nonconvex. In fact, if* $f(\boldsymbol{\beta})$ *can be decomposed as* $l(\boldsymbol{\beta}) + P(\boldsymbol{\beta})$ *with* $0 \leq \boldsymbol{\Delta}_l \leq L\mathbf{D}_2$ *and* $\boldsymbol{\Delta}_P + \mathcal{L}_0\mathbf{D}_2 \geq 0$ *for some*

finite $\mathcal{L}_0 \geq 0$, then setting $\mu_0 = 0$, $\psi_0 = l - \mathcal{L}\|\cdot\|_2^2/2$ with $\mathcal{L} \geq \mathcal{L}_0$ and repeating the previous arguments, we obtain

$$\begin{aligned}\sum_{t=0}^{T} \frac{R_t}{\theta_t^2 \rho_t} \geq & \sum_{t=0}^{T} \frac{1}{\rho_t}(\rho_t \mathbf{\Delta}_\phi + \mathcal{L}\mathbf{D}_2 - L\mathbf{D}_2)(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) \\ & + \sum_{t=0}^{T-1} \frac{1}{\theta_t^2 \rho_t} \mathbf{\Delta}_l(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t+1)}) + \frac{(\mathcal{L} - \mathcal{L}_0)(1-\theta_T)}{\theta_T^3 \rho_T} \mathbf{D}_2(\boldsymbol{\beta}^{(T+1)}, \boldsymbol{\beta}^{(T)}) \\ & + \sum_{t=0}^{T-1} \mathcal{L} \frac{1-\theta_t}{\theta_t \rho_t} \Big\{ \frac{\mathcal{L} - \mathcal{L}_0}{\mathcal{L}} - (1-\theta_t)\theta_{t+1} \frac{\theta_{t+1}}{\theta_t} \Big\} \mathbf{D}_2(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}).\end{aligned} \tag{A.32}$$

With $\rho_t = \rho_0$, it can be shown that $(1-\theta_t)\theta_{t+1}\frac{\theta_{t+1}}{\theta_t}$ achieves the maximum value 0.1612 at $t = 3$ and so $\mathcal{L} \geq 1.2\mathcal{L}_0$ makes the last term nonnegative. (A varying ρ_t may reduce $\mathcal{L}$ further.) Therefore, under $\mathbf{\Delta}_\phi \geq \sigma \mathbf{D}_2$, we can choose any $\rho_0 \geq (L - \mathcal{L})/\sigma$ and $\mathcal{L} \geq 1.2\mathcal{L}_0$ so that for any $\boldsymbol{\beta}$,

$$\begin{aligned}\frac{f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta})}{\theta_T^2} + T \cdot \operatorname*{avg}_{0 \leq t \leq T} \frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t} + \frac{(\mathcal{L} - \mathcal{L}_0)(1-\theta_T)}{\theta_T^3} \mathbf{D}_2(\boldsymbol{\beta}^{(T+1)}, \boldsymbol{\beta}^{(T)}) \\ \leq \rho_0 \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(0)}), \forall T \geq 0.\end{aligned}$$

Hence for some A independent of T, $f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \min_{0 \leq t \leq T} \mathcal{E}_t(\boldsymbol{\beta}) \leq A/T^2$, or $f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \min_{0 \leq t \leq T}\{\mathbf{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \mathbf{\Delta}_{f(\cdot) - \mathbf{\Delta}_{\psi_0}(\cdot, \boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(t+1)})\} \leq A/T^2$ when ϕ is differentiable.

A.15. Proof of Theorem 7. The construction of the new acceleration scheme and the proof are motivated by Proposition 2 of [57], with the use of GBF calculus. First, from Lemma A.2, given any $\boldsymbol{\beta}'_t$,

$$\begin{aligned}& f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}'_t) + (\rho_t \mathbf{D}_2 - \mathbf{\Delta}_{\bar{\psi}_0})(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) + \mathbf{\Delta}_f(\boldsymbol{\beta}'_t, \boldsymbol{\beta}^{(t+1)}) \\ \leq & (\rho_t \mathbf{D}_2 - \mathbf{\Delta}_{\bar{\psi}_0})(\boldsymbol{\beta}'_t, \boldsymbol{\gamma}^{(t)}) - (\rho_t \mathbf{D}_2 - \mathbf{\Delta}_{\mathbf{\Delta}_{\bar{\psi}_0}(\cdot, \boldsymbol{\gamma}^{(t)})})(\boldsymbol{\beta}'_t, \boldsymbol{\beta}^{(t+1)})\end{aligned} \tag{A.33}$$

Let $\boldsymbol{\beta}'_t = \theta_t \boldsymbol{\beta} + (1-\theta_t)\boldsymbol{\beta}^{(t)}$ with θ_t to be determined. Define $h_t(\cdot) = f(\cdot) - \mathbf{\Delta}_{\bar{\psi}_0}(\cdot, \boldsymbol{\gamma}^{(t)})$. By the definition of $\mathbf{C}$,

$$\begin{aligned}-f(\boldsymbol{\beta}'_t) = & \theta_t \mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) - \mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}'_t, \boldsymbol{\gamma}^{(t)}) \\ & - \theta_t f(\boldsymbol{\beta}) - (1-\theta_t) f(\boldsymbol{\beta}^{(t)}) + \mathbf{C}_{h_t}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t).\end{aligned}$$

Plugging the last equality into (A.33) yields

$$\begin{aligned}& f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) - (1-\theta_t)(f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta})) \\ & + (\rho_t \mathbf{D}_2 - \mathbf{\Delta}_{\bar{\psi}_0})(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) + \mathbf{C}_{h_t}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t) \\ & + \theta_t \mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + (1-\theta_t)\mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \\ \leq & (\rho_t \mathbf{D}_2 - \mathbf{\Delta}_{\bar{\psi}_0})(\boldsymbol{\beta}'_t, \boldsymbol{\gamma}^{(t)}) - (\rho_t \mathbf{D}_2 - \mathbf{\Delta}_{\mathbf{\Delta}_{\bar{\psi}_0}(\cdot, \boldsymbol{\gamma}^{(t)})})(\boldsymbol{\beta}'_t, \boldsymbol{\beta}^{(t+1)}) \\ & + \mathbf{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}'_t, \boldsymbol{\gamma}^{(t)}) - \mathbf{\Delta}_f(\boldsymbol{\beta}'_t, \boldsymbol{\beta}^{(t+1)}) \\ = & \rho_t [\mathbf{D}_2(\boldsymbol{\beta}'_t, \boldsymbol{\gamma}^{(t)}) - \mathbf{D}_2(\boldsymbol{\beta}'_t, \boldsymbol{\beta}^{(t+1)})] - \mathbf{\Delta}_{h_t}(\boldsymbol{\beta}'_t, \boldsymbol{\beta}^{(t+1)})\end{aligned}$$

and based on the definition of $\bar{\psi}_0$ and R_t,

$$\begin{aligned}
&f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) + R_t \\
&+ \rho_t \mathbf{D}_2(\boldsymbol{\beta}_t', \boldsymbol{\beta}^{(t+1)}) + \mathbf{C}_{\mu_0 \mathbf{D}_2(\cdot, \boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t) + \mu_0 \mathbf{D}_2(\boldsymbol{\beta}_t', \boldsymbol{\beta}^{(t+1)}) \\
&+ \theta_t \boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \mathbf{C}_{f(\cdot) - \boldsymbol{\Delta}_{\psi_0}(\cdot, \boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t) + \boldsymbol{\Delta}_{f(\cdot) - \boldsymbol{\Delta}_{\psi_0}(\cdot, \boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}_t', \boldsymbol{\beta}^{(t+1)}) \\
\leq &(1 - \theta_t)(f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta})) + \rho_t \mathbf{D}_2(\boldsymbol{\beta}_t', \boldsymbol{\gamma}^{(t)}).
\end{aligned}$$

From Section A.5, $\mathbf{C}_{\mu_0 \mathbf{D}_2(\cdot, \boldsymbol{\gamma}^{(t)})}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t) = \mu_0 \mathbf{C}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}, \theta_t) = \mu_0 \theta_t (1 - \theta_t) \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)})$ and so

$$\begin{aligned}
&f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) + R_t + \theta_t \mathcal{E}_t(\boldsymbol{\beta}) \\
&+ (\rho_t + \mu_0) \mathbf{D}_2(\boldsymbol{\beta}_t', \boldsymbol{\beta}^{(t+1)}) + \mu_0 \theta_t (1 - \theta_t) \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}) \\
\leq &(1 - \theta_t)(f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta})) + \rho_t \mathbf{D}_2(\boldsymbol{\beta}_t', \boldsymbol{\gamma}^{(t)}).
\end{aligned} \tag{A.34}$$

We would like to write $(\rho_t + \mu_0) \mathbf{D}_2(\theta_t \boldsymbol{\beta} + (1 - \theta_t) \boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)}) + \mu_0 \theta_t (1 - \theta_t) \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)})$ into the form of a multiple of $\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t+1)})$ for some $\boldsymbol{\nu}^{(t+1)}$. This can be done by solving the gradient equation with respect to $\boldsymbol{\beta}$:

$$\boldsymbol{\nu}^{(t+1)} = \frac{(\rho_t + \mu_0) \theta_t \boldsymbol{\beta}^{(t+1)} - \rho_t \theta_t (1 - \theta_t) \boldsymbol{\beta}^{(t)}}{\rho_t \theta_t^2 + \mu_0 \theta_t}. \tag{A.35}$$

On the other hand, $\nabla \rho_t \mathbf{D}_2(\theta_t \boldsymbol{\beta} + (1 - \theta_t) \boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) = \mathbf{0}$ gives

$$\boldsymbol{\nu}^{(t)} = \frac{\boldsymbol{\gamma}^{(t)}}{\theta_t} - \frac{1 - \theta_t}{\theta_t} \boldsymbol{\beta}^{(t)}. \tag{A.36}$$

Combining (A.35) and (A.36) results in

$$\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + \frac{\rho_{t-1} \theta_t (1 - \theta_{t-1})}{\rho_{t-1} \theta_{t-1} + \mu_0} (\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}), \tag{A.37}$$

as in (76a). Therefore, (A.34) becomes

$$\begin{aligned}
&f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) + (\theta_t^2 \rho_t + \mu_0 \theta_t) \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t+1)}) + R_t + \theta_t \mathcal{E}_t(\boldsymbol{\beta}) \\
\leq &(1 - \theta_t)(f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta})) + \theta_t^2 \rho_t \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t)}).
\end{aligned} \tag{A.38}$$

Let $\mu_0 = 0$. It follows from (A.38) that

$$\begin{aligned}
&\frac{1}{\theta_t^2 \rho_t} \big[f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) \big] + \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t+1)}) + \frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t \rho_t} + \frac{R_t}{\theta_t^2 \rho_t} \\
\leq &\frac{(1 - \theta_t)}{\theta_t^2 \rho_t} \big[f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}) \big] + \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t)}), \forall t \geq 0.
\end{aligned} \tag{A.39}$$

Under (77b), we have

$$\begin{aligned}
&\frac{1}{\theta_t^2 \rho_t} \big[f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) \big] + \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t+1)}) + \frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t \rho_t} + \frac{R_t}{\theta_t^2 \rho_t} \\
\leq &\frac{1}{\theta_{t-1}^2 \rho_{t-1}} \big[f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}) \big] + \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t)}), \quad \forall t \geq 1.
\end{aligned} \tag{A.40}$$

Summing (A.40) for $t = T, \dots, 1$ and (A.39) for $t = 0$ gives

$$\begin{aligned}&\frac{1}{\theta_T^2\rho_T}[f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta})] + \sum_{t=0}^{T}\Big(\frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t\rho_t} + \frac{R_t}{\theta_t^2\rho_t}\Big)\\ &\le \frac{1-\theta_0}{\theta_0^2\rho_0}\big[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})\big] + \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(0)}) - \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(T+1)}),\end{aligned}$$

and so the first bound noticing that $\boldsymbol{\nu}^{(0)} = \boldsymbol{\gamma}^{(0)} = \boldsymbol{\beta}^{(0)}$.

Moreover, given any $\mu_0 \ge 0$, from (77b), (A.38) implies for any $t \ge 1$,

$$\begin{aligned}&f(\boldsymbol{\beta}^{(t+1)}) - f(\boldsymbol{\beta}) + (\theta_t^2\rho_t + \mu_0\theta_t)\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t+1)}) + R_t + \theta_t\mathcal{E}_t(\boldsymbol{\beta})\\ &\le (1-\theta_t)[f(\boldsymbol{\beta}^{(t)}) - f(\boldsymbol{\beta}) + (\theta_{t-1}^2\rho_{t-1} + \mu_0\theta_{t-1})\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\nu}^{(t)})].\end{aligned} \tag{A.41}$$

Similar to the proof of Theorem 6, a recursive argument using (A.41) and (A.38) gives the second bound.

REMARK A.4. *Compared with the proof of Theorem 6, the proof here needs to perform a finer analysis of $\mathbf{C}_{h_t}$ (the proof of Corollary 2' uses a similar treatment). Otherwise one would get $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + \theta_t(\theta_{t-1}^{-1} - 1)(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)})$ and $\rho_t\theta_t^2/(1-\theta_t) = \theta_{t-1}^2(\rho_{t-1} + \mu_0)$ in place of (76a), (77b), respectively. Following the same proof, we can show that the resultant algorithm does result in a linear rate when $\mu_0 = \mu > 0$, but offers no acceleration ($\theta_0 = 1/\kappa$) in strongly smooth and convex optimization.*

Finally, Remark A.2 still applies. For example, the second conclusion holds when the '$=$' in (77b) is replaced by '$\le$', and it is straightforward to see that μ_0 can be similarly replaced by a sequence of varying μ_t to speed the convergence.

A.16. Statistical accuracy of LLA iterates. In this subsection, assume $f(\boldsymbol{\beta}) = l(\boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta})$, $l(\boldsymbol{\beta}) = l_0(\boldsymbol{X}\boldsymbol{\beta})$, $P(\varrho\boldsymbol{\beta}) = \sum_j P(\varrho\beta_j)$ (by a slight abuse of notation), $P(0) = 0$, $P'_+(0) < +\infty$, $P(t) = P(-t) \ge 0$, $P(t)$ is differentiable for any $t > 0$, and P is concave on $(0, +\infty)$. Recall $\boldsymbol{\Delta}_{\mathrm{LLA}}^{(t)} = \boldsymbol{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1 - P(\cdot)}$ which does *not* satisfy the strong idempotence.

ASSUMPTION $\mathcal{A}(\varepsilon, \delta, \vartheta, K, \boldsymbol{\alpha}, \boldsymbol{\beta})$ Given $\boldsymbol{X}, \boldsymbol{\alpha}, \boldsymbol{\beta}$, there exist $\varepsilon > 0, \delta > 0, \vartheta > 0, K \ge 0$ such that the following inequality holds

$$\begin{aligned}&(1+\varepsilon)\boldsymbol{\Delta}_{\|\boldsymbol{\alpha}\circ(\cdot)\|_1 - P(\cdot)}(\varrho\boldsymbol{\beta}^*, \varrho\boldsymbol{\beta}) + \delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*, \boldsymbol{X}\boldsymbol{\beta}) + \vartheta P_H(\varrho(\boldsymbol{\beta} - \boldsymbol{\beta}^*); \lambda)\\ &\le 2\bar{\boldsymbol{\Delta}}_l(\boldsymbol{\beta}^*, \boldsymbol{\beta}) + P(\varrho\boldsymbol{\beta}; \lambda) - P(\varrho\boldsymbol{\beta}^*; \lambda) + K\lambda^2 J^*.\end{aligned}$$

PROPOSITION 5. *Assume that for any given $T \ge 1$, $\mathcal{A}(\varepsilon, \delta, \vartheta, K, \boldsymbol{\alpha}^{(t)}, \boldsymbol{\beta}^{(t)})$ $(1 \le t \le T)$ is satisfied for some $\varepsilon > 0, \delta > 0, \vartheta > 0, K \ge 0$. Let $\lambda = A\sigma\sqrt{\log(ep)}/\sqrt{(\delta \wedge \vartheta)\vartheta}$. Then the following inequality holds with probability at least $1 - Cp^{-cA^2}$*

$$\boldsymbol{\Delta}_{\mathrm{LLA}}^{(T)}(\varrho\boldsymbol{\beta}^*, \varrho\boldsymbol{\beta}^{(T)}) \le \kappa^T\boldsymbol{\Delta}_{\mathrm{LLA}}^{(0)}(\varrho\boldsymbol{\beta}^*, \varrho\boldsymbol{\beta}^{(0)}) + \frac{\kappa}{1-\kappa}K\lambda^2 J^*,$$

where $\kappa = 1/(1+\varepsilon)$ and C, c are universal positive constants.

PROOF. From the proof of Proposition 4, for any $\boldsymbol{\beta}$,

$$\begin{aligned}&\boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\mathrm{LLA}}^{(t)}(\cdot, \varrho\boldsymbol{\beta}^{(t)})}(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^{(t+1)})\\ &\le f(\boldsymbol{\beta}) - f(\boldsymbol{\beta}^{(t+1)}) + \boldsymbol{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}^{(t)}) - \boldsymbol{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta}^{(t+1)}, \varrho\boldsymbol{\beta}^{(t)}).\end{aligned}$$

Using the definition of $\mathbf{\Delta}_{\mathrm{LLA}}^{(t)}$, we have

$$\begin{aligned}&\mathbf{\Delta}_P(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t+1)})-\mathbf{\Delta}_{\mathbf{\Delta}_{P(\cdot,\varrho\boldsymbol{\beta}^{(t)})}}(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t+1)})\\&+\mathbf{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^{(t+1)})+\sum_j\alpha_j^{(t)}\mathbf{\Delta}_{\mathbf{\Delta}_1(\cdot,\varrho\beta_j^{(t)})}(\varrho\beta_j,\varrho\beta_j^{(t+1)})\\&\le f(\boldsymbol{\beta})-f(\boldsymbol{\beta}^{(t+1)})+\mathbf{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t)}),\end{aligned}\tag{A.42}$$

where we used $\mathbf{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta}^{(t+1)},\varrho\boldsymbol{\beta}^{(t)})\ge 0$ since $P(\cdot)$ is concave on $(0,+\infty)$.

LEMMA A.7. *For any $P(\cdot)$ which is differentiable on $(0,+\infty)$ and satisfies $P(t)=P(-t)\ge 0, P(0)=0$ and $P'_+(0)<+\infty$, we have $\mathbf{\Delta}_{\mathbf{\Delta}_P(\cdot,\alpha)}(\beta,\gamma)=\mathbf{\Delta}_P(\beta,\gamma)-P'_+(0)\mathbf{\Delta}_1(\beta,\gamma)1_{\alpha=0}$ for any $\alpha,\beta,\gamma\in\mathbb{R}$. In particular, $\mathbf{\Delta}_{\mathbf{\Delta}_1(\cdot,\alpha)}(\beta,\gamma)=\mathbf{\Delta}_1(\beta,\gamma)1_{\alpha\ne 0}$.*

The result can be shown from the proof of Lemma 2. Indeed, from (29),

$$\mathbf{\Delta}_P(\cdot,\alpha)=\begin{cases}P(\cdot)-P(\alpha)-P'(\alpha)(\cdot-\alpha), & \alpha\ne 0\\ P(\cdot)-P'_+(0)|\cdot|, & \alpha=0.\end{cases}$$

When $\alpha\ne 0$, by Lemma 1 and Lemma 2, $\mathbf{\Delta}_{\mathbf{\Delta}_P(\cdot,\alpha)}(\beta,\gamma)=\mathbf{\Delta}_P(\beta,\gamma)-\mathbf{\Delta}_{P(\alpha)+P'(\alpha)(\cdot-\alpha)}(\beta,\gamma)=\mathbf{\Delta}_P(\beta,\gamma)$. When $\alpha=0$, $\mathbf{\Delta}_{\mathbf{\Delta}_P(\cdot,\alpha)}(\beta,\gamma)=\mathbf{\Delta}_P(\beta,\gamma)-P'_+(0)\mathbf{\Delta}_1(\beta,\gamma)$. Combining the two cases gives

$$\mathbf{\Delta}_{\mathbf{\Delta}_P(\cdot,\alpha)}(\beta,\gamma)=\mathbf{\Delta}_P(\beta,\gamma)-P'_+(0)\mathbf{\Delta}_1(\beta,\gamma)1_{\alpha=0}.$$

When $P(\boldsymbol{\beta})=\|\boldsymbol{\beta}\|_1$, $\mathbf{\Delta}_{\mathbf{\Delta}_1(\cdot,\alpha)}(\beta,\gamma)=\mathbf{\Delta}_1(\beta,\gamma)-\mathbf{\Delta}_1(\beta,\gamma)1_{\alpha=0}=\mathbf{\Delta}_1(\beta,\gamma)1_{\alpha\ne 0}$.

From Lemma A.7,

$$\mathbf{\Delta}_P(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t+1)})-\mathbf{\Delta}_{\mathbf{\Delta}_{P(\cdot,\varrho\boldsymbol{\beta}^{(t)})}}(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t+1)})=\sum_{j:\beta_j^{(t)}=0}P'_+(0)\mathbf{\Delta}_1(\varrho\beta_j,\varrho\beta_j^{(t+1)})$$

and

$$\sum_j\alpha_j^{(t)}\mathbf{\Delta}_{\mathbf{\Delta}_1(\cdot,\varrho\beta_j^{(t)})}(\varrho\beta_j,\varrho\beta_j^{(t+1)})=\sum_{j:\beta_j^{(t)}\ne 0}\alpha_j^{(t)}\mathbf{\Delta}_1(\varrho\beta_j,\varrho\beta_j^{(t+1)}).$$

Plugging these into (A.42) gives $\mathbf{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^{(t+1)})+\sum_{j:\beta_j^{(t)}=0}P'_+(0)\mathbf{\Delta}_1(\varrho\beta_j,\varrho\beta_j^{(t+1)})+\sum_{j:\beta_j^{(t)}\ne 0}\alpha_j^{(t)}\mathbf{\Delta}_1(\varrho\beta_j,\varrho\beta_j^{(t+1)})\le f(\boldsymbol{\beta})-f(\boldsymbol{\beta}^{(t+1)})+\mathbf{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t)})$.

Together with $\alpha_j^{(t)}=|P'_+(\beta_j^{(t)})|\le P'_+(0)$, we have

$$\mathbf{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}^{(t+1)})+\sum_j\alpha_j^{(t)}\mathbf{\Delta}_1(\varrho\beta_j,\varrho\beta_j^{(t+1)})\le f(\boldsymbol{\beta})-f(\boldsymbol{\beta}^{(t+1)})+\mathbf{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}^{(t)}).\tag{A.43}$$

Letting $\boldsymbol{\beta}=\boldsymbol{\beta}^*$ and using the definition of $\boldsymbol{\epsilon}$, we obtain

$$\begin{aligned}&2\bar{\mathbf{\Delta}}_l(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})+\mathbf{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1}(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t+1)})+P(\varrho\boldsymbol{\beta}^{(t+1)};\lambda)\\&\le\mathbf{\Delta}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t)})+\langle\boldsymbol{\epsilon},\boldsymbol{X}\boldsymbol{\beta}^{(t+1)}-\boldsymbol{X}\boldsymbol{\beta}^*\rangle+P(\varrho\boldsymbol{\beta}^*;\lambda).\end{aligned}$$

From the regularity condition,

$$\begin{aligned}&(1+\varepsilon)\mathbf{\Delta}_{\mathrm{LLA}}^{(t+1)}(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t+1)})+\delta\mathbf{D}_2(\boldsymbol{X}\boldsymbol{\beta}^*,\boldsymbol{X}\boldsymbol{\beta}^{(t+1)})+\vartheta P_H(\varrho(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^*);\lambda)\\&\le 2\bar{\mathbf{\Delta}}_l(\boldsymbol{\beta}^*,\boldsymbol{\beta}^{(t+1)})+\mathbf{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1}(\varrho\boldsymbol{\beta}^*,\varrho\boldsymbol{\beta}^{(t+1)})+P(\varrho\boldsymbol{\beta}^{(t+1)};\lambda)-P(\varrho\boldsymbol{\beta}^*;\lambda)+K\lambda^2J^*\end{aligned}$$

for $1 \le t \le T$. The final conclusion can be proved by combining the last two inequalities and then applying a similar probabilistic argument as in Theorem 5. □

A.17. A-estimators as F-estimators. In this part, we show that an important class of **A**-estimators that has *alternative* optimality, typically arising from block coordinate descent (BCD) algorithms, can often be converted to F-estimators, and analyzed in a similar way. Let $\boldsymbol{\beta} = [\boldsymbol{\beta}_{[1]}^T, \ldots, \boldsymbol{\beta}_{[K]}^T]^T$ where $\boldsymbol{\beta}_{[k]}$ is the kth block, $1 \le k \le K$, and we use $\boldsymbol{\beta}_{[-k]}$ to denote the subvector after removing the kth block. Assume

$$f = l + P$$

where l is differentiable, and P is separable: $P(\boldsymbol{\beta}) = \sum P_k(\boldsymbol{\beta}_{[k]})$. When viewed as a function of $\boldsymbol{\beta}_{[k]}$ only, f is denoted by $f(\boldsymbol{\beta}_{[k]}; \boldsymbol{\beta}_{[-k]})$. We say $\hat{\boldsymbol{\beta}}$ has alternative optimality or is an A-estimator if

$$\hat{\boldsymbol{\beta}}_{[k]} \in \arg\min_{\boldsymbol{\beta}_{[k]}} f(\boldsymbol{\beta}_{[k]}; \hat{\boldsymbol{\beta}}_{[-k]}), 1 \le k \le K. \tag{A.44}$$

LEMMA A.8. *Let $\hat{\beta}$ be an A-estimator of $\min f(\boldsymbol{\beta})$. Construct a surrogate function:*

$$g_{\boldsymbol{\rho}}(\boldsymbol{\beta}; \boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) - \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^-) + \sum \rho_k \mathbf{D}_2(\boldsymbol{\beta}_{[k]}, \boldsymbol{\beta}_{[k]}^-) \tag{A.45}$$

where $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$ with $\rho_k \ge 0$.

(i) If P_k are directionally differentiable and

$$\boldsymbol{\Delta}_{P_k} + \mathcal{L}_k \mathbf{D}_2 \ge 0 \tag{A.46}$$

for some $\mathcal{L}_k \ge 0$, then for any $\rho_k \ge \mathcal{L}_k$, $\hat{\boldsymbol{\beta}}$ must satisfy

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} g_{\boldsymbol{\rho}}(\boldsymbol{\beta}; \boldsymbol{\beta}^-)|_{\boldsymbol{\beta}^- = \hat{\boldsymbol{\beta}}}. \tag{A.47}$$

(ii) If l as a function of $\boldsymbol{\beta}_k$ satisfies $\boldsymbol{\Delta}_{l(\cdot; \boldsymbol{\beta}_{[-k]})} \le L_k \mathbf{D}_2, \forall \boldsymbol{\beta}_{[-k]}$, $1 \le k \le K$, or less restrictively,

$$\boldsymbol{\Delta}_{\hat{l}_k(\cdot)}(\boldsymbol{\beta}_{[k]}, \hat{\boldsymbol{\beta}}_{[k]}) \le L_k \mathbf{D}_2(\boldsymbol{\beta}_{[k]}, \hat{\boldsymbol{\beta}}_{[k]}), \ \forall \boldsymbol{\beta}_{[k]} \tag{A.48}$$

where $\hat{l}_k(\boldsymbol{\beta}_{[k]})$ denotes $l_k(\boldsymbol{\beta}_{[k]}; \hat{\boldsymbol{\beta}}_{[-k]})$, then for any $\rho_k > L_k$, (A.47) still holds. In addition, if $\hat{\boldsymbol{\beta}}_{[k]}$ is the unique solution to (A.44), then $\rho_k \ge L_k$ suffices.

Overall, (A.47) provides a useful *joint* optimization form that can be used as the so-called "basic inequality" in empirical process theory, and so with the lemma, A-estimators can analyzed like F-estimators. Moreover, the quality of the initial point can be incorporated in the analysis; see [54].

PROOF. (i) The condition (A.46) means that $g_{\boldsymbol{\rho}}$ is convex in, $\boldsymbol{\beta}_{[k]}$. By Lemma 4 and Lemma 1, and the fact that $g_{\boldsymbol{\rho}}$ is separable in $\boldsymbol{\beta}_{[k]}$, $1 \le k \le K$, we immediately know that $\hat{\boldsymbol{\beta}}$ is necessarily a solution to $\min_{\boldsymbol{\beta}} g_{\boldsymbol{\rho}}(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}})$.

(ii) We use a shorthand notation $\hat{g}_k(\boldsymbol{\beta}_{[k]})$ to denote $g_{\boldsymbol{\rho}}(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}})$ as a function of $\boldsymbol{\beta}_{[k]}$ when $\boldsymbol{\beta}_{[-k]} = \hat{\boldsymbol{\beta}}_{[-k]}$. Let $\tilde{\boldsymbol{\beta}}_{[k]} \in \arg\min \hat{g}_k(\boldsymbol{\beta}_{[k]})$. It suffices to show $\tilde{\boldsymbol{\beta}}_{[k]} = \hat{\boldsymbol{\beta}}_{[k]}$. Because of the separability of $g_{\boldsymbol{\rho}}$,

$$f(\tilde{\boldsymbol{\beta}}_{[k]}; \hat{\boldsymbol{\beta}}_{[-k]}) + (\rho_k - L_k)\mathbf{D}_2(\hat{\boldsymbol{\beta}}_{[k]}, \tilde{\boldsymbol{\beta}}_{[k]}) \le f(\hat{\boldsymbol{\beta}}_{[k]}; \hat{\boldsymbol{\beta}}_{[-k]})$$

and so $(\rho_k - L_k)\mathbf{D}_2(\hat{\boldsymbol{\beta}}_{[k]}, \tilde{\boldsymbol{\beta}}_{[k]}) = 0$. The conclusion follows. □

Some conclusions like (i) can be extended to functions defined on Riemannian manifolds. It is also worth mentioning that in the regression setup, which is of primary interest in many statistical applications, we can use some surrogates with $\rho_k = 1$, regardless of the design or penalty, to convert alternative optimality to joint optimality. The following lemma exemplifies the point in matrix regression, and is condition free.

LEMMA A.9. *Let $l_0(\boldsymbol{A};\boldsymbol{Y}) = \|\boldsymbol{Y}-\boldsymbol{A}\|_F^2/2$, and $\boldsymbol{A}$ be defined differently as follows.*
(i) Let $\boldsymbol{A} = \sum \boldsymbol{X}_k\boldsymbol{B}_k$ with $\boldsymbol{B} = (\boldsymbol{B}_1,\dots,\boldsymbol{B}_K)$, where the dependence of $\boldsymbol{B}$ (and $\boldsymbol{X}_k$) is dropped for simplicity. Consider the problem

$$\min_{\boldsymbol{B}_1,\dots,\boldsymbol{B}_K} l_0(\boldsymbol{A};\boldsymbol{Y}) + \sum P_k(\boldsymbol{B}_k) \text{ s.t. } \boldsymbol{A} = \sum \boldsymbol{X}_k\boldsymbol{B}_k. \tag{A.49}$$

Then the set of A-estimators of (A.49) *is exactly the set of F-estimators associated with the following surrogate*

$$g(\boldsymbol{B},\boldsymbol{B}^-) = l_0(\boldsymbol{A};\boldsymbol{Y}) - \mathbf{D}_{l_0}(\boldsymbol{A},\boldsymbol{A}^-) + \sum P_k(\boldsymbol{B}_k) + \sum \mathbf{D}_2(\boldsymbol{X}_k\boldsymbol{B}_k,\boldsymbol{X}_k\boldsymbol{B}_k^-). \tag{A.50}$$

(ii) Let $\boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}_1\cdots\boldsymbol{B}_K$ with $\boldsymbol{B} = (\boldsymbol{B}_1,\dots,\boldsymbol{B}_K)$, where the dependence of $\boldsymbol{B}$ (and $\boldsymbol{X}_k$) is dropped for simplicity. Consider

$$\min_{\boldsymbol{B}_1,\dots,\boldsymbol{B}_K} l_0(\boldsymbol{A};\boldsymbol{Y}) + \sum P_k(\boldsymbol{B}_k) \text{ s.t. } \boldsymbol{A} = \boldsymbol{X}\boldsymbol{B}_1\cdots\boldsymbol{B}_K. \tag{A.51}$$

Redefine l_0 as a function $\bar{l}_0$ of $\boldsymbol{B}$ and introduce a discrepancy measure d^2 as follows

$$l_0(\boldsymbol{A};\boldsymbol{Y}) = \bar{l}_0(\boldsymbol{B};\boldsymbol{X},\boldsymbol{Y})$$

$$d^2(\boldsymbol{B},\boldsymbol{B}^-) = \frac{1}{2}\sum_{k=1}^K \|\boldsymbol{X}\boldsymbol{B}_1^-\cdots\boldsymbol{B}_{k-1}^-(\boldsymbol{B}_k-\boldsymbol{B}_k^-)\boldsymbol{B}_{k+1}^-\cdots\boldsymbol{B}_K^-\|_F^2.$$

Then the set of A-estimators of (A.51) *is exactly the set of F-estimators associated with the surrogate*

$$\begin{aligned} g(\boldsymbol{B},\boldsymbol{B}^-) &= \bar{l}_0(\boldsymbol{B}) - \boldsymbol{\Delta}_{\bar{l}_0}(\boldsymbol{B},\boldsymbol{B}^-) + \sum P_k(\boldsymbol{B}_k) + d^2(\boldsymbol{B},\boldsymbol{B}^-) \\ &= \frac{1}{2}\|\boldsymbol{A}^- - \boldsymbol{Y}\|_F^2 + \langle \boldsymbol{X}\boldsymbol{B}_1^-\cdots\boldsymbol{B}_K^- - \boldsymbol{Y}, \\ &\quad \sum_{k=1}^K \boldsymbol{X}\boldsymbol{B}_1^-\cdots\boldsymbol{B}_{k-1}^-(\boldsymbol{B}_k-\boldsymbol{B}_k^-)\boldsymbol{B}_{k+1}^-\cdots\boldsymbol{B}_K^-\rangle + \sum P_k(\boldsymbol{B}_k) \\ &\quad \frac{1}{2}\sum_{k=1}^K \|\boldsymbol{X}\boldsymbol{B}_1^-\cdots\boldsymbol{B}_{k-1}^-(\boldsymbol{B}_k-\boldsymbol{B}_k^-)\boldsymbol{B}_{k+1}^-\cdots\boldsymbol{B}_K^-\|_F^2. \end{aligned} \tag{A.52}$$

The lemma can be directly proved by the definition of GBF and matrix differentiation and its proof is omitted. For the application of the first result (i), see [50] for example. The second result can be used to study bilinear problems or NMF like matrix decomposition problems. One could show a statistical accuracy result in terms of d^1 (which satisfies $d^1 \le Kd^2$),

$$d^1(\boldsymbol{B},\boldsymbol{B}^-) = \frac{1}{2}\Big\|\sum_{k=1}^K \boldsymbol{X}\boldsymbol{B}_1^-\cdots\boldsymbol{B}_{k-1}^-(\boldsymbol{B}_k-\boldsymbol{B}_k^-)\boldsymbol{B}_{k+1}^-\cdots\boldsymbol{B}_K^-\Big\|_F^2,$$

under a proper regularity condition involving d^2; see, for example, [53].

A.18. Statistical error analysis of a general optimal solution. This part demonstrates that using the statistical notions and Bregman calculus developed earlier can perform statistical analysis of a general optimization problem that may not be in the MLE setup:

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) \text{ s.t. } \boldsymbol{\beta} \in \mathcal{S} \tag{A.53}$$

where f is directionally differentiable and $\mathcal{S} \subset \mathbb{R}^p$ can be formulated by linear equality constraints $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\alpha}$, sparsity constraints $\|\boldsymbol{\beta}\|_0 \le s$, nonnegativity constraints $\boldsymbol{\beta} \ge 0$, and so on.

Statistically, we would like to study how a target parameter can be recovered from solving (A.53) in the present of data noise. Following (53), let $\boldsymbol{\beta}^*$ be a statistical truth and define the associated effective noise by $\boldsymbol{\epsilon} = -\nabla f(\boldsymbol{\beta}^*)$, assuming f is differentiable at $\boldsymbol{\beta}^*$.

Although $\boldsymbol{\beta}^*$ in the above definition can be any point, a meaningful recovery must be under some conditions satisfied the associated $\boldsymbol{\epsilon}$. Consider the following three scenarios:

(a) Statistical estimation often assumes a zero mean noise:

$$\mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}, \tag{A.54}$$

which essentially means that the statistical truth makes the gradient of the *expectation* of f (so as to remove data randomness) vanish—see Section 3.2. Yet (A.54) alone does not always guarantee a unique $\boldsymbol{\beta}^*$.

(b) Stronger conclusions can be obtained for the $\boldsymbol{\beta}^*$ that satisfies the no-model-ambiguity assumption: f is differentiable at $\boldsymbol{\beta}^* \in D = \text{dom}(f)$ with the gradient $\nabla f(\boldsymbol{\beta}^*) = -\boldsymbol{\epsilon}$, $\boldsymbol{\beta}^*$ is a finite optimal solution to the Fenchel conjugate as $\boldsymbol{\zeta} = -\boldsymbol{\epsilon}$:

$$f^*(\boldsymbol{\zeta}) = \sup_{\boldsymbol{\beta}} \langle \boldsymbol{\zeta}, \boldsymbol{\beta} \rangle - f(\boldsymbol{\beta}), \tag{A.55}$$

and the extended real-valued convex function f^* is differentiable at $-\boldsymbol{\epsilon}$. This assumption simply means that $(\boldsymbol{\beta}^*, -\boldsymbol{\epsilon})$ makes a so-called "conjugate pair". Note that f need not be overall strictly convex, especially when D is compact according to Danskin's min-max theorem [6].

(c) Another popular assumption in statistical learning is strong convexity in a restricted sense (especially when $\boldsymbol{\beta} = \boldsymbol{A}\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_0 \le s$):

$$(\boldsymbol{\Delta}_f - \mu \mathbf{D}_2)(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) \ge 0, \ \forall \boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathcal{S} \tag{A.56}$$

for some $\mu > 0$. The condition may hold even when the number of unknowns is much larger than the sample size [12, 36].

The following theorem uses the GBF calculus to argue how the statistical accuracy of the obtained solutions is determined by the (tail decay of) effective noise. Probabilistic arguments can follow to bound the stochastic terms more explicitly.

THEOREM A.1. *Let $\hat{\boldsymbol{\beta}}$ be an optimal solution to* (A.53).

(i) Under the zero mean assumption (A.54) *and $\boldsymbol{\beta}^* \in \mathcal{S}$, the risk of $\hat{\boldsymbol{\beta}}$ in terms of $\boldsymbol{\Delta}_f$ satisfies a Fenchel-Young form bound*

$$\mathbb{E}\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \mathbb{E}[f^*(\boldsymbol{\epsilon}) + f(\boldsymbol{\beta}^*)]. \tag{A.57}$$

(ii) Under the no-model-ambiguity assumption in (b) and $\boldsymbol{\beta}^* \in \mathcal{S}$, we have

$$\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \boldsymbol{\Delta}_{f^*}(\boldsymbol{\epsilon}, -\boldsymbol{\epsilon}). \tag{A.58}$$

(iii) An oracle inequality holds for any $\delta > 0$ and any reference $\boldsymbol{\beta} \in \mathcal{S}$:

$$(\boldsymbol{\Delta}_f - \delta \mathbf{D}_2)(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \frac{1}{2\delta}[\sup_{\boldsymbol{\theta} \in \Gamma(\boldsymbol{\beta})} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle]^2, \tag{A.59}$$

where $\Gamma(\boldsymbol{\beta}) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_2 \le 1, \boldsymbol{\theta} = \bar{\boldsymbol{\beta}} - \boldsymbol{\beta}$ for some $\bar{\boldsymbol{\beta}} \in \mathcal{S}\}$. In particular, under (A.56),

$$\mathbf{D}_2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \frac{1}{\mu}\boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \frac{1}{2\mu^2}[\sup_{\boldsymbol{\theta} \in \Gamma(\boldsymbol{\beta})} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle]^2. \tag{A.60}$$

The first two bounds reveal the important role of the Fenchel conjugate of the loss, and can be made more explicit under proper Orlicz norm conditions of $\boldsymbol{\epsilon}$; the third conclusion, on the basis of the supremum of an empirical process [60], demonstrates how modern probabilistic tools can be used to derive finite-sample error bounds of $\hat{\boldsymbol{\beta}}$ in a general noisy setup.

PROOF. First, by definition, $f(\hat{\boldsymbol{\beta}}) \le f(\boldsymbol{\beta}^*)$, from which it follows that $\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \langle \boldsymbol{\epsilon}, \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \rangle$. Define

$$h(\boldsymbol{\delta}) = \boldsymbol{\Delta}_f(\boldsymbol{\delta} + \boldsymbol{\beta}^*, \boldsymbol{\beta}^*).$$

By assumption, f is a proper function and applying Fenchel-Young's inequality gives

$$\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \langle \boldsymbol{\epsilon}, \boldsymbol{\delta} \rangle|_{\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*} \le \frac{1}{c} \boldsymbol{\Delta}_f(\boldsymbol{\delta} + \boldsymbol{\beta}^*, \boldsymbol{\beta}^*)|_{\boldsymbol{\delta} = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*} + \frac{1}{c} h^*(c\boldsymbol{\epsilon})$$

or $(1 - 1/c)\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le h^*(c\boldsymbol{\epsilon})/c$ for any $c > 0$.

On the other hand,

$$\begin{aligned} h^*(\boldsymbol{\zeta}) &= \sup_{\boldsymbol{\delta}} \langle \boldsymbol{\zeta}, \boldsymbol{\delta} \rangle - f(\boldsymbol{\beta}^* + \boldsymbol{\delta}) + f(\boldsymbol{\beta}^*) + \langle \nabla f(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle \\ &= \sup_{\boldsymbol{\delta}} \langle \boldsymbol{\zeta} + \nabla f(\boldsymbol{\beta}^*), \boldsymbol{\delta} \rangle - f(\boldsymbol{\beta}^* + \boldsymbol{\delta}) + f(\boldsymbol{\beta}^*) \\ &= \sup_{\boldsymbol{\delta}} \langle \boldsymbol{\zeta} + \nabla f(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* + \boldsymbol{\delta} \rangle - f(\boldsymbol{\beta}^* + \boldsymbol{\delta}) + f(\boldsymbol{\beta}^*) - \langle \boldsymbol{\zeta} + \nabla f(\boldsymbol{\beta}^*), \boldsymbol{\eta}^* \rangle \\ &= f^*(\boldsymbol{\zeta} + \nabla f(\boldsymbol{\beta}^*)) + f(\boldsymbol{\beta}^*) - \langle \boldsymbol{\zeta} + \nabla f(\boldsymbol{\beta}^*), \boldsymbol{\beta}^* \rangle, \end{aligned}$$

where $f(\boldsymbol{\beta}^*), \nabla f(\boldsymbol{\beta}^*)$ are known to be finite. Therefore we obtain

$$\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \frac{1}{c-1} [f^*((c-1)\boldsymbol{\epsilon}) + f(\boldsymbol{\beta}^*) - (c-1)\langle \boldsymbol{\epsilon}, \boldsymbol{\beta}^* \rangle], \ \forall c > 0. \tag{A.61}$$

Taking $c = 2$ and using $\mathbb{E}\boldsymbol{\epsilon} = \mathbf{0}$ gives the $\boldsymbol{\Delta}_f$ risk bound (A.57).

Next, we prove the second bound under the no-model-ambiguity assumption. Using the optimality of $\boldsymbol{\beta}^*$, we have further

$$h^*(\boldsymbol{\zeta}) = f^*(\boldsymbol{\zeta} + \nabla f(\boldsymbol{\beta}^*)) - f^*(-\boldsymbol{\epsilon}) - \langle \boldsymbol{\zeta}, \boldsymbol{\beta}^* \rangle.$$

Moreover, from the assumption and definition (A.55), it is easy to show that $\boldsymbol{\beta}^* \in \partial f^*(-\boldsymbol{\epsilon})$, and so $\nabla f^*(-\boldsymbol{\epsilon}) = \boldsymbol{\beta}^*$, from which it follows that

$$h^*(\boldsymbol{\zeta}) = \boldsymbol{\Delta}_{f^*}(\boldsymbol{\zeta} - \boldsymbol{\epsilon}, -\boldsymbol{\epsilon}). \tag{A.62}$$

Taking $c = 2$ gives (A.58) (even though $\mathbb{E}\boldsymbol{\epsilon}$ may not be $\mathbf{0}$).

Finally, for any $\boldsymbol{\beta} \in \mathcal{S}$, $f(\hat{\boldsymbol{\beta}}) \le f(\boldsymbol{\beta})$ and so

$$\boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \langle \boldsymbol{\epsilon}, (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})/\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \rangle \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2.$$

We obtain a general result

$$(\boldsymbol{\Delta}_f - \delta \mathbf{D}_2)(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \frac{1}{2\delta} [\sup_{\boldsymbol{\theta} \in \Gamma(\boldsymbol{\beta})} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle]^2, \tag{A.63}$$

for any $\delta > 0$.

Based on the regularity condition,

$$\frac{\delta}{2} \mathbf{D}_2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \boldsymbol{\Delta}_f(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) - \frac{\delta}{2} \mathbf{D}_2(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \boldsymbol{\Delta}_f(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \frac{1}{\delta} (\sup_{\boldsymbol{\theta} \in \Gamma(\boldsymbol{\beta})} \langle \boldsymbol{\epsilon}, \boldsymbol{\theta} \rangle)^2$$

for any $\delta \le \mu$. Taking $\delta = \mu$ gives the desired result. □

Algorithm B.1 Accelerated Bregman of the second kind

Input $\boldsymbol{\beta}^{(0)}$: initial value; $\rho_{\min} > 0$, $\alpha > 0$, $M \in \mathbb{N}$, $\mu_0 \geq 0$ (e.g., $\rho_{\min} = 1$, $\alpha = 2$, $M = 3$)

$\theta_0 \in (0, 1]$, $t \leftarrow 0$, $\boldsymbol{\alpha}^{(0)} \leftarrow \boldsymbol{\beta}^{(0)}$;
while not converged **do**
  $\rho_t \leftarrow \rho_{\min}/\alpha$, $s \leftarrow 0$
  repeat
    $s \leftarrow s + 1$
    $\rho_t \leftarrow \alpha\rho_t$
    if $t \geq 1$, then $\theta_t \leftarrow (\sqrt{r^2 + 4r} - r)/2$ with $r = (\rho_{t-1}\theta_{t-1} + \mu_0)\theta_{t-1}/\rho_t$
    $\boldsymbol{\gamma}^{(t)} \leftarrow (1 - \theta_t)\boldsymbol{\beta}^{(t)} + \theta_t\boldsymbol{\alpha}^{(t)}$
    $\boldsymbol{\alpha}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\beta}}\{f(\boldsymbol{\beta}) - \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \mu_0\boldsymbol{\Delta}_{\phi}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \theta_t\rho_t\boldsymbol{\Delta}_{\phi}(\boldsymbol{\beta}, \boldsymbol{\alpha}^{(t)})\}$
    $\boldsymbol{\beta}^{(t+1)} \leftarrow (1 - \theta_t)\boldsymbol{\beta}^{(t)} + \theta_t\boldsymbol{\alpha}^{(t+1)}$
    $R_t \leftarrow \theta_t^2\rho_t\boldsymbol{\Delta}_{\phi}(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\alpha}^{(t)}) - \boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)})$
      $+(1 - \theta_t)\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) + \mathbf{C}_{f(\cdot) - \boldsymbol{\Delta}_{\bar{\psi}_0}(\cdot, \boldsymbol{\gamma}^{(t)})}(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t)}, \theta_t)$
  until $R_t \geq 0$ or $s > M$
  if $s > M$, pick $(\boldsymbol{\alpha}^{(t+1)}, \boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}, \rho_t, \theta_t)$ with the largest $R_t/(\theta_t^2\rho_t)$
  $t \leftarrow t + 1$
end while
return $\boldsymbol{\beta}^{(t+1)}$.

APPENDIX B: ALGORITHMS FOR ACCELERATIONS

For clarity, we give an outline of the algorithms for acceleration.

Algorithm B.2 Accelerated Bregman of the first kind

Input $\boldsymbol{\beta}^{(0)}$: initial value; $\rho_{\min} > 0$, $\alpha > 0$, $M \in \mathbb{N}$, $\mu_0 \geq 0$ ($\rho_{\min} = 1$, $\alpha = 2$, $M = 3$)

$\theta_0 \in (0, 1]$, $t \leftarrow 0$;
while not converged **do**
  $\rho_t \leftarrow \rho_{\min}/\alpha$, $s \leftarrow 0$
  repeat
    $s \leftarrow s + 1$
    $\rho_t \leftarrow \alpha\rho_t$
    if $t \geq 1$, then $\theta_t \leftarrow (\sqrt{r^2 + 4r} - r)/2$ with $r = (\rho_{t-1}\theta_{t-1} + \mu_0)\theta_{t-1}/\rho_t$
    $\boldsymbol{\gamma}^{(t)} \leftarrow \boldsymbol{\beta}^{(t)} + \{\rho_{t-1}\theta_t(1 - \theta_{t-1})/(\rho_{t-1}\theta_{t-1} + \mu_0)\}(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)})$
      if $t \geq 1$ and $\boldsymbol{\beta}^{(t)}$ if $t = 0$
    $\boldsymbol{\beta}^{(t+1)} \leftarrow \arg\min_{\boldsymbol{\beta}}\{f(\boldsymbol{\beta}) - \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \mu_0\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)}) + \rho_t\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)})\}$
    $R_t \leftarrow (\rho_t\mathbf{D}_2 - \boldsymbol{\Delta}_{\bar{\psi}_0})(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}) + (1 - \theta_t)\boldsymbol{\Delta}_{\bar{\psi}_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)})$
  until $R_t \geq 0$ or $s > M$
  if $s > M$, pick $(\boldsymbol{\beta}^{(t+1)}, \boldsymbol{\gamma}^{(t)}, \rho_t, \theta_t)$ with the largest associated $R_t/(\theta_t^2\rho_t)$
  $t \leftarrow t + 1$
end while
return $\boldsymbol{\beta}^{(t+1)}$.

APPENDIX C: EXPERIMENTS

This section performs some simulation studies to support the theoretical results.

C.1. Computational error. In this part, we use mirror descent and DC programming to solve two nonconvex problems.

• *Nonconvex mirror descent for IS divergence minimization.* In infrared astronomical satellite (IRAS) image reconstruction [13] and audio signal processing [21, 35], the Itakura-Saito (IS) divergence (or the negative cross Burg entropy), $\mathrm{IS}(\boldsymbol{a},\boldsymbol{b}) = \sum_i (a_i/b_i - \log(a_i/b_i) - 1)$, is popularly used to measure the discrepancy between the observed data and the reconstructed data. Given $\boldsymbol{X} \in \mathbb{R}_+^{n\times p}$, and $\boldsymbol{y} \in \mathbb{R}_+^n$, the problem can be defined by $\min f(\boldsymbol{\beta}) := \mathrm{IS}(\boldsymbol{y}, \boldsymbol{X}\boldsymbol{\beta})$ s.t. $\boldsymbol{\beta} \in \mathbb{R}_+^p$, which is nonconvex in $\boldsymbol{\beta}$. To maintain the nonnegativity constraint in updating $\boldsymbol{\beta}$ automatically, we develop a mirror descent algorithm. Concretely, define $g(\boldsymbol{\beta};\boldsymbol{\beta}^-) = f(\boldsymbol{\beta}) + (\rho\mathbf{D}_\varphi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta},\boldsymbol{\beta}^-)$, where $\varphi(\boldsymbol{\beta}) = \sum_j \beta_j \log \beta_j - \beta_j$. Then, minimizing $g(\boldsymbol{\beta};\boldsymbol{\beta}^{(t)})$ with respect to $\boldsymbol{\beta}$ gives rise to a *multiplicative* rule

$$\beta_j^{(t+1)} = \beta_j^{(t)} \exp\Big[-\frac{1}{\rho} \sum_i \frac{(\boldsymbol{X}\boldsymbol{\beta}^{(t)})_i - y_i}{(\boldsymbol{X}\boldsymbol{\beta}^{(t)})_i^2} X_{ij} \Big]. \tag{C.1}$$

From Theorem 2, the $\mathcal{O}(1/T)$ rate of convergence holds for the optimization error $\mathrm{avg}_{0\le t\le T}(2\rho\bar{\mathbf{D}}_\varphi - \boldsymbol{\Delta}_f)(\boldsymbol{\beta}^{(t+1)},\boldsymbol{\beta}^{(t)})$. To verify this, we generated a design matrix of size 1000×1000 with all elements drawn from $U(0,1)$, and set $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{e}$ with β_j^* chosen uniformly from the interval $(0,5)$ and $e_i \sim \mathcal{N}(0,\sigma^2)$ with $\sigma^2 = 10$. We fixed $1/\rho = 0.01$. Figure C.1 shows how the logarithm of optimization error converges to $-\infty$ (since $\log 0 = -\infty$) for 50 different $\boldsymbol{\beta}^{(0)}$ with $\beta_j^{(0)}$ randomly chosen from $U(0,1)$. Observe that all the error curves in the log-log plot are bounded by a line with slope -1.

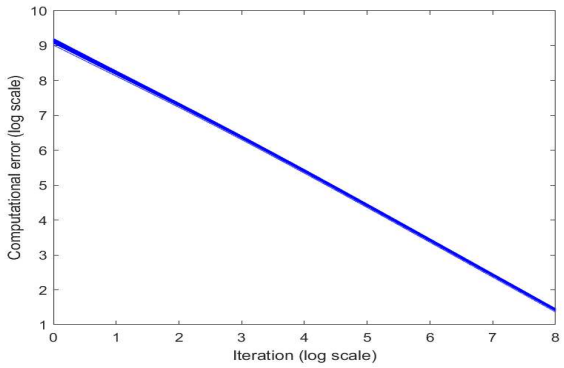

FIG C.1. *Log-log plot of optimization error v.s. number of iterations: mirror descent for IS divergence minimization with 50 random starting points. All error curves are bounded above by the dashed line which has slope* -1.

• *DC programming for capped-ℓ_1 SVM.* High-dimensional classification with concurrent feature selection can be achieved by minimizing a composite objective function. Given $\boldsymbol{X} = [\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n]^\top \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \{-1,1\}^n$, let $l(\boldsymbol{\beta}) = \sum_{i=1}^n (1 - y_i \boldsymbol{x}_i^\top \boldsymbol{\beta})_+$ be the hinge loss [61] that is nondifferentiable, and $P(\boldsymbol{\beta};\lambda) = \sum_{j=1}^p \min(\lambda|\beta_j|, \lambda^2/2)$ be the capped-ℓ_1 penalty [66] which is nonsmooth and nonconvex. [42] proposed an effective DC algorithm for solving $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) := l(\boldsymbol{\beta}) + P(\boldsymbol{\beta};\lambda)$ based on the decomposition $P(\boldsymbol{\beta};\lambda) = d_1(\boldsymbol{\beta};\lambda) - d_2(\boldsymbol{\beta};\lambda)$ with

$$d_1(\boldsymbol{\beta};\lambda) = \lambda\|\boldsymbol{\beta}\|_1, \quad d_2(\boldsymbol{\beta};\lambda) = \sum_{j=1}^p \max(\lambda|\beta_j| - \lambda^2/2, 0). \tag{C.2}$$

As stated in Example 4, we can recharacterize DC as a Bregman-surrogate algorithm

$$\text{(C.3a)} \qquad \boldsymbol{\beta}^{(t+1)} \in \arg\min f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{d_2}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)})$$

$$\text{(C.3b)} \qquad \in \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \max(0, 1 - y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \big(|\beta_j| - \beta_j 1_{|\beta_j^{(t)}| \geq \lambda/2} \big).$$

(C.3b) is equivalent to a linear program: $\min_{\boldsymbol{\xi},\boldsymbol{\zeta},\boldsymbol{\beta}} \sum_{i=1}^{n} \xi_i + \lambda \sum_{j=1}^{p} \zeta_j - \lambda \sum_{j=1}^{p} \beta_j 1_{|\beta_j^{(t)}| \geq \lambda/2}$ s.t. $\xi_i \geq 1 - y_i \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for $i \in [n]$, $-\zeta_j \leq \beta_j \leq \zeta_j$ for $j \in [p]$ and $\xi_i, \zeta_j \geq 0$ for $i \in [n], j \in [p]$, which can be efficiently solved by standard linear programming (LP) solvers. For the convergence of the DC algorithm, a similar result can be shown for the optimization error $\text{avg}_{0 \leq t \leq T}(\boldsymbol{\Delta}_l + \hat{\boldsymbol{\Delta}}_{d_2} + \boldsymbol{\Delta}_{d_1})(\boldsymbol{\beta}^{(t)}, \boldsymbol{\beta}^{(t+1)})$, following the lines of the proof of Proposition 4.

We generated $\boldsymbol{X} \in \mathbb{R}^{400 \times 800}$ with each row following $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $\Sigma_{ij} = 0.5^{|i-j|}$, $\boldsymbol{\beta}^* = [15, 10, 0, \ldots, 0]^\top$, and $\boldsymbol{y} = \text{sgn}(\boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{e})$ where $e_i \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 10$. We fixed $\lambda = 1$ and ran the DC algorithm for 50 different starting points with each component randomly drawn from $U(0, 1)$. The corresponding optimization error curves are plotted in Figure C.2, where the $\mathcal{O}(1/T)$ rate of convergence is impressive.

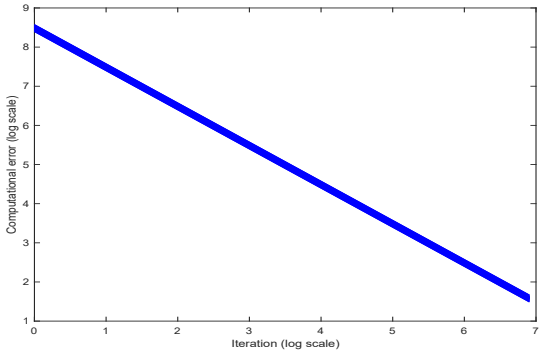

FIG C.2. *Log-log plot of optimization error v.s. number of iterations: DC programming for capped-ℓ_1 SVM (50 different initial points).*

C.2. Statistical error. In this part, we consider two algorithms for sparse regression: LLA and iterative thresholding. The nonconvex "hard" penalty defined by (55) is applied, which is constructed from the hard-thresholding rule via (48). Given $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$, we study the following regularized problem $\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) := l(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_H(\varrho\beta_j; \lambda)$, where l is the loss function and $\varrho = \|\boldsymbol{X}\|_2$. The loss functions under consideration are the ordinary quadratic loss and a nonconvex loss which is resistant to gross outliers:

(i) ℓ_2 loss: $l(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2$;

(ii) Tukey's biweight loss: $l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{|\boldsymbol{x}_i^\top \boldsymbol{\beta} - y_i|} \psi(t)\mathrm{d}t$ and $\psi(t) = t[1 - (t/c)^2]^2$ if $|t| \leq c$ and 0 otherwise, where $c = 4.685\sigma$ with σ a robust estimate of the standard deviation of errors [26].

In either case, we have a nonconvex optimization problem. In simulations, the design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has i.i.d. rows drawn from $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ with $\Sigma_{ij} = 0.15^{|i-j|}$, the response is given by $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{e}$ with $e_i \sim N(0, \sigma^2)$ and $\sigma^2 = 10$, and the regularization parameter λ is set to $A\sigma\sqrt{\log(ep)}$. We set $n = 800, p = 1000, \boldsymbol{\beta}^* = [12, 8, 0, \ldots, 0]^\top$ and $A = 2$.

First, we tested the statistical accuracy of LLA (cf. Theorem 3 and Proposition 5). We generated 15 initial points $\boldsymbol{\beta}^{(0)}$ with each element following $U(-a,a)$, with $a \in \{0.5, 1, 1.5\}$ and 5 for each. Figure C.3 shows how the statistical error varies as the cycles progress, with each curve representing an average over 20 implementations of the same setting. Here, the errors are plotted on a log scale for a better view of the convergence rate. Unlike Figure C.1 and Figure C.2, the statistical errors can not reach 0 (or $-\infty$ in the log plot) due to the existence of noise. But they all achieved essentially the same order of statistical precision, which verifies Theorem 3, and the statistical convergence of LLA was really fast.

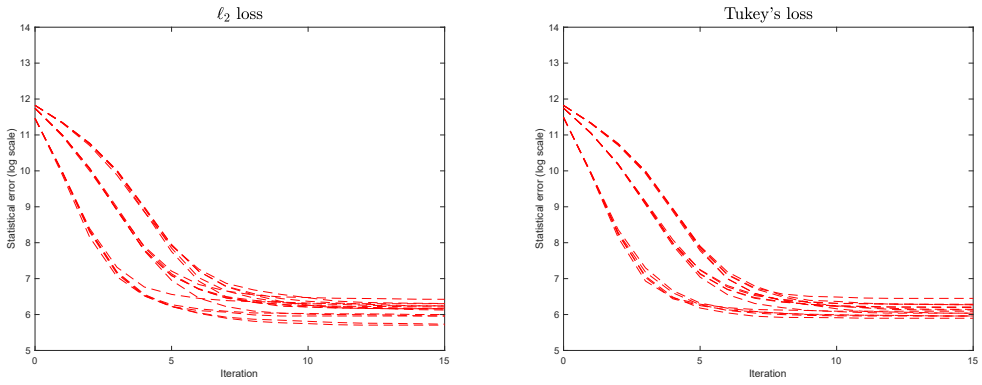

FIG C.3. *Log plot of the statistical accuracy of LLA iterates in P_H-penalized sparse regression (left) and robust regression (right).*

On the other hand, the computational burden of LLA turned out to be pretty high, mainly due to the cost of solving a weighted lasso problem at each iteration. We thus turned to iterative thresholding because of its low per-iteration complexity. Figure C.4 shows some analogous results. According to Figure C.4, all final statistical errors were controlled within the same order of precision. The convergence process seems to conform to the bound in Theorem 5: when t is small, $\log \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}) \lesssim -\log(1/\kappa)\, t + \log(\boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(0)}))$, and when t is large, $\log \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(t)}) \lesssim \kappa^t \boldsymbol{\Delta}_\psi(\boldsymbol{\beta}^*, \boldsymbol{\beta}^{(0)}) + \log(\kappa K \lambda^2 J^*/(1-\kappa))$, demonstrating an exponential decay.

C.3. Accelerations. We test the acceleration schemes in IS divergence minimization and robust sparse regression in this subsection.

Figure C.5 shows the power of applying the (second) acceleration in IS divergence minimization problem in Section C.1, where we used 50 starting points with $\beta_j^{(0)} \sim U(0,1), 1 \le j \le p$. With the acceleration, the number of iterations was brought down from 1000 to less than 50 to reach the same value of the objective function, and the overall computational time was saved by nearly 90%.

Figure C.6 shows the convergence of statistical error when applying the (first) acceleration scheme in iterative thresholding for the P_H-penalized Tukey's loss minimization problem as mentioned in Section C.2. The simulation setting remains the same as before and we sampled 20 initial points with $\beta_j^{(0)} \sim U(-1,1), 1 \le j \le p$. A substantial reduction in the number of iterations was achieved. Of course, the line search causes some overhead in computation. But the accelerated iterative thresholding still reduced the overall running time by more than 30%, and obtained slightly better statistical accuracy.

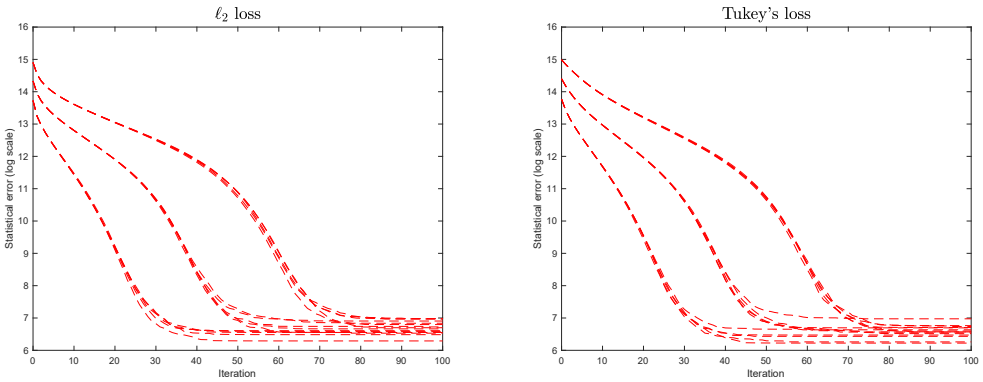

FIG C.4. *Log plot of the statistical accuracy of iterative thresholding iterates in P_H-penalized sparse regression (left) and robust regression (right).*

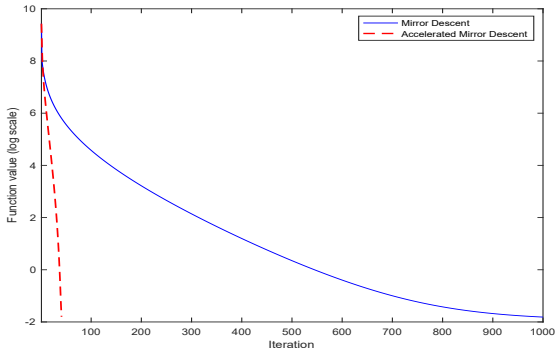

FIG C.5. *Objective function value (shown on* log *scale) v.s. number of iterations for the plain and accelerated exponentiated gradient descent algorithms in nonconvex Burg entropy optimization.*

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics* **40** 2452–2482.

[2] AN, L. T. H. and TAO, P. D. (2005). The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems. *Annals of Operations Research* **133** 23–46.

[3] BANERJEE, A., MERUGU, S., DHILLON, I. S. and GHOSH, J. (2005). Clustering with Bregman Divergences. *Journal of Machine Learning Research (JMLR)* **6**.

[4] BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences* **2** 183–202.

[5] BEN-TAL, A. and NEMIROVSKI, A. (2013). Optimization III: Convex Analysis, Nonlinear Programming Theory, Standard Nonlinear Programming Algorithms. Lecture Notes.

[6] BERTSEKAS, D. P. (1999). *Nonlinear Programming*, 2nd ed. Athena Scientific.

[7] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *The Annals of Statistics* 1705–1732.

[8] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis* **27** 265–274.

[9] BREGMAN, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics* **7** 200–217.

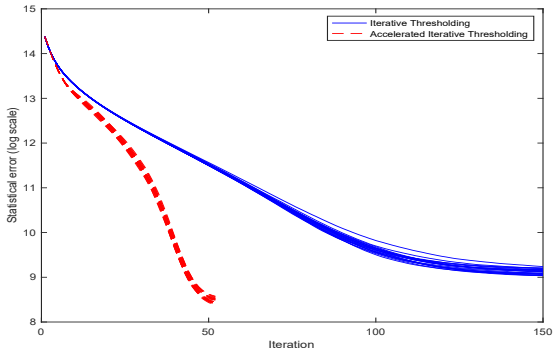

FIG C.6. *Statistical error (shown on* log *scale) v.s. iteration number for iterative thresholding and accelerated iterative thresholding in robust sparse regression.*

[10] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the LASSO. *Electronic Journal of Statistics* **1** 169–194.

[11] CANDES, E. and TAO, T. (2007). The Dantzig selector: Satistical estimation when p is much larger than n. *The Annals of Statistics* 2313–2351.

[12] CANDES, E. J. and TAO, T. (2005). Decoding by Linear Programming. *IEEE Trans. Inf. Theor.* **51** 4203–4215.

[13] CAO, EGGERMONT, P. P. B. and TEREBEY, S. (1999). Cross Burg entropy maximization and its application to ringing suppression in image reconstruction. *IEEE Transactions on Image Processing* **8** 286–292.

[14] CHEN, G. and TEBOULLE, M. (1993). Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions. *SIAM Journal on Optimization* **3** 538–543.

[15] CICHOCKI, A., ICHI AMARI, S., ZDUNEK, R., KOMPASS, R., HORI, G. and HE, Z. (2006). Extended SMART algorithms for non-negative matrix factorization. In *ICAISC* (L. RUTKOWSKI, R. TADEUSIEWICZ, L. A. ZADEH and J. M. ZURADA, eds.). *Lecture Notes in Computer Science* **4029** 548-562. Springer.

[16] CORREA, R., JOFRE, A. and THIBAULT, T. (1994). Subdifferential Monotonicity as Characterization of Convex Functions. *Numerical Functional Analysis and Optimization* **15** 531–535.

[17] DONOHO, D. and JOHNSTONE, I. (1994). Ideal Spatial Adaptation via Wavelet Shrinkages. *Biometrika* **81** 425–455.

[18] DUCHI, J. C., HAZAN, E. and SINGER, Y. (2011). Adaptive Subgradient Methods for Online Learning and Stochastic Optimization. *J. Mach. Learn. Res.* **12** 2121-2159.

[19] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96** 1348–1360.

[20] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *The Annals of Statistics* **42** 819–849.

[21] FÉVOTTE, C., BERTIN, N. and DURRIEU, J.-L. (2009). Nonnegative Matrix Factorization with the Itakura-saito Divergence: With Application to Music Analysis. *Neural Computation* **21** 793–830.

[22] FRANK, I. E. and FRIEDMAN, J. H. (1993). A Statistical View of Some Chemometrics Regression Tools. *Technometrics* **35** 109-135.

[23] GASSO, G., RAKOTOMAMONJY, A. and CANU, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Transactions on Signal Processing* **57** 4686-4698.

[24] GHADIMI, S. and LAN, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Mathematical Programming* **156** 59–99.

[25] HAGOOD, J. W. and THOMSON, B. S. (2006). Recovering a function from a Dini derivative. *The American Mathematical Monthly* **113** 34-46.

[26] HAMPEL, F. R., RONCHETTI, E. M., ROUSSEEUW, P. J. and STAHEL, W. A. (2005). *Robust Statistics.* John Wiley & Sons, New York.

[27] HUBER, P. J. (1981). *Robust Statistics*. John Wiley and Sons, New York.

[28] HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *The American Statistician* 30–37.

[29] HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *The Annals of Statistics* **33** 1617–1642.

[30] JØRGENSEN, B. (1987). Exponential Dispersion Models. *Journal of the Royal Statistical Society. Series B* **49** 127-145.
[31] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer.
[32] KRICHENE, W., BAYEN, A. and BARTLETT, P. L. (2015). Accelerated Mirror Descent in Continuous and Discrete Time. In *Advances in Neural Information Processing Systems* (C. CORTES, N. LAWRENCE, D. LEE, M. SUGIYAMA and R. GARNETT, eds.) **28**. Curran Associates, Inc.
[33] LANGE, K. and ZHOU, H. (2014). MM algorithms for geometric and signomial programming. *Mathematical Programming* **143** 339–356.
[34] LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature* **401** 788–791.
[35] LEFÈVRE, A., BACH, F. R. and FÉVOTTE, C. (2011). Online algorithms for nonnegative matrix factorization with the Itakura-Saito divergence. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, WASPAA 2011* 313–316.
[36] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: statistical and algorithmic theory for local optima. *Journal Machine Learning Research* **16** 559–616.
[37] LOUNICI, K., PONTIL, M., TSYBAKOV, A. B. and VAN DE GEER, S. (2011). Oracle Inequalities and Optimal Inference under Group Sparsity. *Annals of Statistics* **39** 2164-2204.
[38] NEMIROVSKI, A. and YUDIN, D. (1983). *Problem complexity and method efficiency in optimization*. *Wiley-Interscience series in discrete mathematics*. Wiley, Chichester, New York.
[39] NESTEROV, Y. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady* **27** 372–376.
[40] NESTEROV, Y. (1988). On an approach to the construction of optimal methods of minimization of smooth convex functions. *Ekonom. i. Mat. Metody (In Russian)* **24** 509–517.
[41] NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. *Applied optimization*. Kluwer Academic Publ., Boston, Dordrecht, London.
[42] ONG, C. S. and AN, L. T. H. (2013). Learning sparse classifiers with difference of convex functions algorithms. *Optimization Methods and Software* **28** 830-854.
[43] PAN, W., SHEN, X. and LIU, B. (2013). Cluster analysis: unsupervised learning via supervised learning with a non-convex penalty. *Journal of Machine Learning Research* **14** 1865–1889.
[44] RIGOLLET, P. and TSYBAKOV, A. (2011). Exponential Screening and optimal rates of sparse estimation. *Annals of Statistics* **39** 731–771.
[45] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. Princeton University Press, Princeton, NJ.
[46] SCHMIDT, M. (2010). Graphical Model Structure Learning with ℓ_1-Regularization, PhD thesis, University of British Columbia.
[47] SHE, Y. (2009). Thresholding-based Iterative Selection Procedures for Model Selection and Shrinkage. *Electronic Journal of Statistics* **3** 384-415.
[48] SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Computational Statistics and Data Analysis* **9** 2976-2990.
[49] SHE, Y. (2016). On the finite-sample analysis of Θ-estimators. *Electronic Journal of Statistics* **10** 1874–1895.
[50] SHE, Y. and CHEN, K. (2017). Robust reduced-rank regression. *Biometrika* **104** 633-647.
[51] SHE, Y., HE, Y. and WU, D. (2014). Learning Topology and Dynamics of Large Recurrent Neural Networks. *IEEE Transactions on Signal Processing* **62** 5881-5891.
[52] SHE, Y., LI, H., WANG, J. and WU, D. (2013). Grouped Iterative Spectrum Thresholding for Super-Resolution Sparse Spectrum Selection. *IEEE Transactions on Signal Processing* **61** 6371-6386.
[53] SHE, Y., SHEN, J. and ZHANG, C. Supervised Multivariate Learning with Simultaneous Feature Autogrouping and Dimension Reduction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. To appear.
[54] SHE, Y., WANG, Z. and SHEN, J. Gaining Outlier Resistance with Progressive Quantiles: Fast Algorithms and Theoretical Studies. *Journal of the American Statistical Association*. To appear.
[55] TAO, P. D. and SOUAD, E. B. (1986). Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. *North-Holland Mathematics Studies* **129** 249-271.
[56] TIBSHIRANI, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society (Series B)* **58** 267-288.
[57] TSENG, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. Technical report, Department of Mathematics, University of Washington.
[58] TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer, New York, NY.
[59] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* **3** 1360–1392.

[60] VAN DER VAART, A. W. and WELLNER, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer.
[61] VAPNIK, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, NY, USA.
[62] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning* **1** 1-305.
[63] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *The Annals of Statistics* **42** 2164–2201.
[64] ZHANG, C., JIANG, Y. and CHAI, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika* **97** 551–566.
[65] ZHANG, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38** 894-942.
[66] ZHANG, T. (2010). Analysis of Multi-stage Convex Relaxation for Sparse Regularization. *Journal of Machine Learning Research* **11** 1081–1107.
[67] ZOU, H. and LI, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *Annals of Statistics* **36** 1509–1533.