The Annals of Statistics
2021, Vol. 49, No. 6, 3434–3459
https://doi.org/10.1214/21-AOS2090

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, [*]*yshe@stat.fsu.edu;* [†]*wzfmath@gmail.com;* [‡]*jj17g@my.fsu.edu*

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 [16, 18, 53], 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 [8] 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 [52]. The generalized Bregman surrogate framework has a close connection

Received February 2021; revised May 2021.

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

Key words and phrases. Nonconvex optimization, nonsmooth optimization, MM algorithms, Bregman divergence, statistical algorithmic analysis, momentum-based acceleration.

to the majorization-minimization (MM) principle [22, 23]. 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 $g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) \ge f(\boldsymbol{\beta})$; the benefits will be seen in step size control and acceleration.

A variety of algorithms can be recharacterized by Bregman surrogates, including DC programming [44], local linear approximation (LLA) [55] and iterative thresholding [7, 38]. 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, 17, 51]; in particular, [29], 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 tth 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 over-optimization 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 [33, 34]. Accelerated gradient algorithms [4, 26, 46] 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 [20] 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 (in the Supplementary Material [43]).

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 $\langle \boldsymbol{A}, \boldsymbol{B}\rangle = \mathrm{tr}\{\boldsymbol{A}^\top \boldsymbol{B}\}$. If $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite; we also write $\boldsymbol{A} \succeq \boldsymbol{B}$. Let $[p] := \{1, \dots, p\}$. 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 [36]. 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 [8], 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 $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathrm{dom}(\psi)$ 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 $\boldsymbol{\Delta}_{\psi} \ge \boldsymbol{\Delta}_{\phi}$ to denote $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma}) \ge \boldsymbol{\Delta}_{\phi}(\boldsymbol{\beta},\boldsymbol{\gamma})$ for all $\boldsymbol{\beta},\boldsymbol{\gamma}$, and so $\boldsymbol{\Delta}_{\psi} \ge 0$ stands for $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma}) \ge 0, \forall \boldsymbol{\beta},\boldsymbol{\gamma}$. 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 $\boldsymbol{\Delta}_{\psi} \ge 0$; conversely, if ψ is directionally differentiable and $\boldsymbol{\Delta}_{\psi} \ge 0$ 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) $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma}) = \int_0^1 [\delta\psi(\boldsymbol{\gamma}+t(\boldsymbol{\beta}-\boldsymbol{\gamma}); \boldsymbol{\beta}-\boldsymbol{\gamma}) - \delta\psi(\boldsymbol{\gamma};\boldsymbol{\beta}-\boldsymbol{\gamma})]\,\mathrm{d}t$, *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 $\boldsymbol{\Delta}_f \ge -\nu\mathbf{D}_2$, 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: $\|\nabla\psi(\boldsymbol{\beta}) - \nabla\psi(\boldsymbol{\gamma})\|_* \le L\|\boldsymbol{\beta}-\boldsymbol{\gamma}\|$ for any $\boldsymbol{\beta},\boldsymbol{\gamma}$, where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$, $\boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma}) \le L\|\boldsymbol{\beta}-\boldsymbol{\gamma}\|^2/2$ and for the Euclidean norm, $\boldsymbol{\Delta}_{\psi} \le L\mathbf{D}_2$ 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, $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) \le \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma})$, and when ψ is concave, $\boldsymbol{\Delta}_{\boldsymbol{\Delta}_{\psi}(\cdot,\boldsymbol{\alpha})}(\boldsymbol{\beta},\boldsymbol{\gamma}) \ge \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta},\boldsymbol{\gamma})$ 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 $\mathbf{D}_\psi(\boldsymbol{\beta},\boldsymbol{\gamma}) = \mathbf{D}_\psi(\boldsymbol{\beta},\boldsymbol{\alpha}) + \mathbf{D}_\psi(\boldsymbol{\alpha},\boldsymbol{\gamma}) - \langle \boldsymbol{\beta}-\boldsymbol{\alpha}, \nabla\psi(\boldsymbol{\gamma}) - \nabla\psi(\boldsymbol{\alpha})\rangle$ [12]. 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 [24]. 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 [50], 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 $\boldsymbol{a} = \mathbf{0}$. 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 [50].

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*

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

for all $\boldsymbol{\eta} \in \Omega$, $\boldsymbol{z} \in \mathrm{ri}(\mathcal{M})$, 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 $\mathrm{ri}(\mathcal{M})$ becomes $\mathcal{M}^\circ$.*

(ii) *As long as Ω is open,*

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

for all $\boldsymbol{\eta} \in \Omega$, $\boldsymbol{z} \in \mathrm{ri}(\mathcal{M})$. If $p_{\boldsymbol{\eta}}$ is also minimal, $\mathbf{\Delta}_{b^*} = \mathbf{D}_{b^*}$ and $\mathrm{ri}(\mathcal{M}) = \mathcal{M}^\circ$.

(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*

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

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 $[y_1,\dots,y_m] \in \mathcal{Y} = \{y_k \in \{0,1\}, 1 \le k \le m, \sum y_k = 1\}$ $(n = 1)$, $\mathbb{E}y_k \propto \exp(\eta_k)$ 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 $[\mu_1,\ldots,\mu_m] \in \mathcal{M} = \{[\mu_k] : \sum\mu_k = 1, \mu_k \ge 0\}$ and $+\infty$ otherwise. Clearly, b^* is not differentiable (given any $z \in \mathrm{ri}(\mathcal{M})$, $\partial b^*(z) = \{\log z + t\mathbf{1} : t \in \mathbb{R}\}$), but nicely our two GBF representations still hold. In addition, if the right-hand side of (10) or (11), as a function of z, is continuous on $\overline{\mathcal{M}}$, which is the case for Bernoulli, multinomial and Poisson, (i) and (ii) hold for any $z \in \overline{\mathcal{M}}$ from [50], 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 [47]. More concretely, consider a general signal class

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

where $s^* \le p$, $0 \le M \le +\infty$. 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 $I(t) = 1_{t \ge c}$. 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* $p \ge 2$, $1 \le s^* \le p/2$. *Let*

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

(i) *If*

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

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

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

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

(ii) *If*

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

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

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

The GBF-form conditions (15), (16) can be viewed as an extension of restricted isometry [10], 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 [31]: $\boldsymbol{\beta}^{(t+1)} = (\nabla \varphi)^{-1}(\nabla \varphi(\boldsymbol{\beta}^{(t)}) - \nabla f(\boldsymbol{\beta}^{(t)})/\rho)$, 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 [45]), bridge penalties [18], SCAD [16], capped-ℓ_1 [54] and MCP [53]. There is a universal connection between thresholding rules and penalty functions [39], 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 $\min_{\boldsymbol{\beta}} l(\boldsymbol{\beta}) + \sum_j P(\varrho \beta_j; \lambda)$ [7, 38]:

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

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 [38] 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 $s\in\mathbb{R}$ [40].

A closely related *iterative quantile-thresholding* procedure [39, 42] 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) [28] 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 [13]:

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

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

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:

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

leads to the multiplicative update formulas.

EXAMPLE 4 (DC programming). DC programming [44] is capable of tackling a large class of nonsmooth nonconvex optimization problems; see, for example, [19, 35]. 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

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

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 $\boldsymbol{\gamma}^{(t)}=\delta d_2(\boldsymbol{\beta}^{(t)};\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)})\times(\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)})/\|\boldsymbol{\beta}^{(t+1)}-\boldsymbol{\beta}^{(t)}\|_2^2$ 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 [55] 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

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

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

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

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 [41].

Let $\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\dots,\boldsymbol{x}_n]^\top\in\mathbb{R}^{n\times p}$ be the data matrix, and $\boldsymbol{y}=[y_1,\dots,y_n]^\top$ 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 $\nabla^2 f(\boldsymbol{\beta}) \preceq \boldsymbol{X}^\top \operatorname{diag}\{|0.1y_i| + 0.08\}_{i=1}^n \boldsymbol{X}$, which motivates a Bregman surrogate

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

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 $\boldsymbol{B} = \boldsymbol{X}^\top \operatorname{diag}\{|0.1y_i| + 0.08\}_{i=1}^n\boldsymbol{X}$, $\boldsymbol{u}^{(t)} = \pi(\boldsymbol{X}^\top\boldsymbol{\beta}^{(t)})$ and $\circ$ denotes the Hadamard product. This type of surrogate functions is closely related to proximal Newton-type methods [37] and signomial programming [27].

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

$$(31) \qquad \boldsymbol{\beta}^{(t+1)} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^{(t)}) := f(\boldsymbol{\beta}) + \boldsymbol{\Delta}_{\psi}(\boldsymbol{\beta}, \boldsymbol{\beta}^{(t)}), \quad t \ge 0$$

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 $g(\boldsymbol{\beta}; \boldsymbol{\beta}^-) \ge f(\boldsymbol{\beta})$ 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,

$$(32) \qquad \left\{\boldsymbol{\beta} : \boldsymbol{\beta} \in \operatorname*{arg\,min}_{\boldsymbol{\beta}} g(\boldsymbol{\beta}; \boldsymbol{\beta}^-)\Big|_{\boldsymbol{\beta}^- = \boldsymbol{\beta}}\right\},$$

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

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

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

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

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

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

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

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 [29].)

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*

$$2\bar{\mathbf{\Delta}}_f \geq \varepsilon \mathbf{\Delta}_\psi \tag{39}$$

for some $\varepsilon > 0$, *then*

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

for any $T \geq 0$.

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

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

The $\mathbf{D}_2$-form bound is classical for problems with strong convexity; see, for example, Theorem 2.1.15 in [32]. 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

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

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 $\mathbf{\Delta}_\psi$ 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

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

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

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

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* $T \geq 1$,

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

(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 $\bar{\mathbf{\Delta}}_{\psi}$ on the left-hand side of (45). This gives a relaxed stepsize control than MM. We use mirror descent $\mathbf{\Delta}_{\psi} = \rho\mathbf{D}_{\varphi} - \mathbf{\Delta}_{f}$ 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 $\mathbf{\Delta}_f \le L\bar{\mathbf{D}}_{\varphi}$ 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 $\mathbf{\Delta}_{\psi} \ge 0$, 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 [15, 46]. 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*

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

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* $T \ge 1$

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

When the loss satisfies $\boldsymbol{\Delta}_l \le L\mathbf{D}_2$, 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, $P(0) = 0$, $P'_+(0) < +\infty$, $P(t) = P(-t) \ge 0$ and $P(t)$ is differentiable for any $t > 0$. Recall its Bregman form surrogate

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

where $\boldsymbol{\alpha}^{(t)} = [\alpha_j^{(t)}]$ with $\alpha_j^{(t)} = |P'_+(\beta_j^{(t)})|$, $1 \le j \le p$. We abbreviate $\boldsymbol{\Delta}_{\|\boldsymbol{\alpha}^{(t)}\circ(\cdot)\|_1 - P(\cdot)}$ to $\boldsymbol{\Delta}_{\mathrm{LLA}}^{(t)}$, which does not satisfy strong idempotence. By combining $\bar{\boldsymbol{\Delta}}_{\mathrm{LLA}}^{(t)}$ 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* $T \ge 1$:

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

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 [55], $2\bar{\boldsymbol{\Delta}}_{\mathrm{LLA}}^{(t)}(\varrho\boldsymbol{\beta},\varrho\boldsymbol{\beta}') + \boldsymbol{\Delta}_f(\boldsymbol{\beta},\boldsymbol{\beta}') = \boldsymbol{\Delta}_l(\boldsymbol{\beta},\boldsymbol{\beta}') + \boldsymbol{\Delta}_{-P}(\varrho\boldsymbol{\beta}',\varrho\boldsymbol{\beta}) + 2\sum_j \alpha_j^{(t)}\bar{\boldsymbol{\Delta}}_1(\varrho\beta_j,\varrho\beta'_j) \ge 0, \forall \boldsymbol{\beta},\boldsymbol{\beta}'$. 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 tth 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 $l(\boldsymbol{\beta}) = l_0(\boldsymbol{\eta}) = l_0(\boldsymbol{X\beta})$ is directionally differentiable and $P_\Theta(\cdot; \lambda)$ is induced by a thresholding Θ via (48). Since l_0 is placed on $\boldsymbol{X\beta}$, 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 $\boldsymbol{X\beta}^*$ (but not necessarily differentiable on all of $\mathbb{R}^n$) and define the *effective noise* by

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

(An alternative assumption is that $\delta l_0(\boldsymbol{X\beta}^*; \boldsymbol{h})$ 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 $\nabla\{\mathbb{E}[l_0(\boldsymbol{X\beta}^*)]\} = 0$ 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 $l(\boldsymbol{\beta}) = l_0(\boldsymbol{X\beta}) = -\langle \boldsymbol{y}, \boldsymbol{X\beta}\rangle + \langle \boldsymbol{1}, b(\boldsymbol{X\beta})\rangle$ (cf. (9) with $\sigma = 1$), and so

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

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 $\boldsymbol{y} = \boldsymbol{X\beta}^* + \boldsymbol{\epsilon}^{\mathrm{raw}}$, $l(\boldsymbol{\beta}) = l_{\mathrm{Huber}}(\boldsymbol{r}) = \sum_{i:|r_i|\leq a\sigma} r_i^2/2 + \sum_{i:|r_i|>a\sigma}(a|r_i| - a^2\sigma^2/2)$ with $\boldsymbol{r} = \boldsymbol{y} - \boldsymbol{X\beta}$ [21], simple calculation gives $\epsilon_i = \epsilon_i^{\mathrm{raw}} 1_{|\epsilon_i^{\mathrm{raw}}|\leq a\sigma} + a\sigma 1_{|\epsilon_i^{\mathrm{raw}}|>a\sigma}$, 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 $\Theta_H(t; \lambda) = t 1_{|t|>\lambda}$ 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 [49], 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 $\boldsymbol{\beta} \in \mathbb{R}^p$:*

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

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* [14, 25] 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

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

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*

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

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 $C > 0$,

$$\begin{aligned}\mathbb{E}\boldsymbol{\Delta}_l(\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^*) \le \mathbb{E}\biggl\{\Bigl(\frac{1+r}{1-r}\Bigr)^2 \boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}^*) + \frac{(1+r)KA^2}{(1-r)(2\vartheta \wedge \delta)\vartheta}\sigma^2 J(\boldsymbol{\beta})\log(ep)\biggr\} \\ + \frac{C(1+r)}{(1-r)(2\vartheta \wedge \delta)}\sigma^2.\end{aligned} \tag{61}$$

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 [25]. 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 $\boldsymbol{\Delta}_{l_0}(\boldsymbol{X\beta}, \boldsymbol{X\beta}^*)$ 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 $P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \vartheta P_H(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \mathcal{L}_\Theta\mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \delta\|\boldsymbol{X\gamma}\|_2^2/2 \le (2 - \varepsilon)\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + K\lambda^2 J + P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda) - \vartheta P_H(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda)$, or $(1 + \vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}}; \lambda) + \mathcal{L}_\Theta\mathbf{D}_2(\varrho\boldsymbol{\beta}, \varrho\boldsymbol{\beta}') + \delta\|\boldsymbol{X\gamma}\|_2^2/2 \le (2 - \varepsilon)\boldsymbol{\Delta}_l(\boldsymbol{\beta}, \boldsymbol{\beta}') + K\lambda^2 J + (1 - \vartheta)P_\Theta(\varrho\boldsymbol{\gamma}_{\mathcal{J}^c}; \lambda)$ since $P_H \le P_\Theta$.

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

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

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)

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

for some $K \ge 0$, $\vartheta > 0$. Taking $\varrho = c\|\boldsymbol{X}\|_2$ results in scale invariance with respect to $\boldsymbol{X}$. Let us compare (63) with the restricted eigenvalue (RE) condition and the compatibility condition [6, 48]. For given $\mathcal{J}$, the two conditions assume that there exist positive numbers κ, ϑ_{RE} such that $J\|\boldsymbol{X\gamma}\|_2^2 \ge \kappa\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1^2$ (compatibility) or more restrictively, $\|\boldsymbol{X\gamma}\|_2^2 \ge \kappa\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2$ (RE), for all $\boldsymbol{\gamma} : (1 + \vartheta_{\mathrm{RE}})\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \geq \|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1$. Therefore, $(1 + \vartheta)\varrho\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \leq K\sqrt{J}\|\boldsymbol{X}\boldsymbol{\gamma}\|_2 \vee \varrho\|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1$ with $K = (1 + \vartheta_{\mathrm{RE}})/(\varrho\sqrt{\kappa})$, $\vartheta = \vartheta_{\mathrm{RE}}$. 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), [29], 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 $\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \geq \alpha\|\boldsymbol{\gamma}\|_2^2 - \tau \log p\|\boldsymbol{\gamma}\|_1^2/n$ for some constant $\alpha > 0$ and $\tau \geq 0$, from which it follows that for any $\boldsymbol{\gamma} : (1 + \vartheta_{\mathrm{RE}})\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1 \geq \|\boldsymbol{\gamma}_{\mathcal{J}^c}\|_1$, $\|\boldsymbol{X}\boldsymbol{\gamma}\|_2^2 \geq \alpha\|\boldsymbol{\gamma}\|_2^2 - \tau(2+\vartheta_{\mathrm{RE}})^2 \frac{\log p}{n}\|\boldsymbol{\gamma}_{\mathcal{J}}\|_1^2 \geq \alpha\|\boldsymbol{\gamma}\|_2^2 - \tau(2+\vartheta_{\mathrm{RE}})^2 \frac{J \log p}{n}\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2 \geq \kappa'\|\boldsymbol{\gamma}_{\mathcal{J}}\|_2^2$, where $\kappa' = \alpha - \tau(2 + \vartheta_{\mathrm{RE}})^2(J \log p/n)$. 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 [29] is that the latter enforces an ℓ_1-type side constraint, for example, $\|\boldsymbol{\beta}\|_1 \leq R$, 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: $\langle \boldsymbol{\epsilon}, \boldsymbol{X}\boldsymbol{\beta}\rangle \leq \|\boldsymbol{X}^\top\boldsymbol{\epsilon}\|_\infty\|\boldsymbol{\beta}\|_1$ (see, e.g., [6, 9, 30]). 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 [29], 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 ($q : 0 \leq q < 1$) [18] 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 [40], 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 [29], 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: $\|\boldsymbol{X}^\top\boldsymbol{y}\|_\infty \leq \lambda$, 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 tth 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 $K \ge 0$,

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

and ψ is differentiable for simplicity. Recall that (39) in Proposition 2 requires $2\bar{\mathbf{\Delta}}_f$ to dominate $\varepsilon\mathbf{\Delta}_{\psi}$; (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*

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

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 [29] 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 [29], 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 1 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 [4, 26, 32, 46]. This section attempts to extend Nesterov's *first* and *second* accelerations [33, 34] 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 tth iteration, with R_t appropriately defined via generalized Bregman notation. We consider the following two broad scenarios to devise the acceleration schemes.

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

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

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

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

where $0 \le \theta \le 1$. Like $\boldsymbol{\Delta}$, $\mathbf{C}$ is a linear operator of ψ and its nonnegativity means convexity. Some connections between $\boldsymbol{\Delta}$ 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 $\mu_0 \ge 0$, $\theta_t \in (0,1]$, $\rho_t > 0$ ($\forall t \ge 0$), 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:

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

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

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* $T \geq 0$,

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

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

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

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 [34, 46], 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

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

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 $\mathcal{L} \ge 0$. Given any $s \le p$ and $\boldsymbol{X}$, define the restricted isometry number $\rho_+(s)$ [11] that satisfies $\|\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \le \rho_+(s)\|\boldsymbol{\beta}\|_2^2$, $\forall \boldsymbol{\beta} : \|\boldsymbol{\beta}\|_0 \le s$, 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 $\mathcal{L} \ge \rho_+(2q)/\sqrt{r}$, 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 $l(\boldsymbol{\beta}^{(T+1)}) - l(\boldsymbol{\beta}^*) + \min_{0 \le t \le T} \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^*, \boldsymbol{\gamma}^{(t)}) \le A/T^2$ for all $T \ge 0$, 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 $\boldsymbol{\Delta}_l \le L\mathbf{D}_2$ and a function $P(\boldsymbol{\beta})$ that can be lifted: $\boldsymbol{\Delta}_P + \mathcal{L}_0\mathbf{D}_2 \ge 0$ for some finite $\mathcal{L}_0 \ge 0$, then one can utilize a ψ_0 as $l - 0.6\mathcal{L}_0\|\cdot\|_2^2$ and a universal ρ_t to fulfill $T \cdot \mathrm{avg}_{t \le T} R_t/(\theta_t^2\rho_t) \ge 0$ 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 $\rho_t/\rho_{t-1} \ge 1 - (at + ab + 1)/(t + b - 1)^2$ ($t \ge 1$) for some constants a, b: $a > -2$, $b \ge a + 1$, induction based on (69b) gives $\theta_t \le (a+2)/(t+b)$ and $\sum_{t=0}^T \rho_T/(\rho_t\theta_t) \ge (T + c_1)^2/(a+2)^2 + c_2$ (with constants c_i dependent on a, b) for any $t \ge 1$, from which it follows that

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

Now, under $R_t \ge 0$ or just $\sum_{t=0}^T R_t/(\theta_t^2\rho_t) \ge 0$, (70) gives $f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \min_{0 \le t \le T} \mathcal{E}_t(\boldsymbol{\beta}) \le \mathcal{O}(\rho_T/T^2)$ 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 $\mu_0 \ge 0$, $\theta_t \in (0, 1]$, $\rho_t > 0$ for all $t \ge 0$, 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

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

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

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* $\rho_t > 0$ $(t \ge 0)$, *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*

$$\frac{f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta})}{\theta_T^2\rho_T} + T\cdot \operatorname*{avg}_{0\le t\le T}\frac{\mathcal{E}_t(\boldsymbol{\beta})}{\theta_t\rho_t} + T\cdot \operatorname*{avg}_{0\le t\le T}\frac{R_t}{\theta_t^2\rho_t}$$
$$\le \mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)}) + \frac{1-\theta_0}{\theta_0^2\rho_0}[f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})] \quad \textit{for any } \boldsymbol{\beta} \textit{ and } T \ge 0.$$

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

$$f(\boldsymbol{\beta}^{(T+1)}) - f(\boldsymbol{\beta}) + \theta_T^2\left(\rho_T + \frac{\mu_0}{\theta_T}\right)\mathbf{D}_2(\boldsymbol{\beta}, (\boldsymbol{\gamma}^{(T+1)} - (1-\theta_{T+1})\boldsymbol{\beta}^{(T+1)})/\theta_{T+1})$$
$$+ \sum_{t=0}^{T}\left(\prod_{s=t+1}^{T}(1-\theta_s)\right)(R_t + \theta_t\mathcal{E}_t(\boldsymbol{\beta}))$$
$$\le \left(\prod_{t=1}^{T}(1-\theta_t)\right)[(1-\theta_0)(f(\boldsymbol{\beta}^{(0)}) - f(\boldsymbol{\beta})) + \theta_0^2\rho_0\mathbf{D}_2(\boldsymbol{\beta}, \boldsymbol{\beta}^{(0)})].$$

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 $(\sqrt{4\kappa - 3} - 1)/(\sqrt{4\kappa - 3} + 1)(\le 1 - \sqrt{3/(4\kappa)})$. (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 [32].

Theorem 7 accommodates diverse choices of the parameters ψ_0, μ_0, ρ_t, θ_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), $\mathcal{E}_t(\boldsymbol{\beta}) \ge \boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}, \boldsymbol{\gamma}^{(t)})$. 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 $\boldsymbol{\Delta}_{\psi_0}(\boldsymbol{\beta}^{(t)}, \boldsymbol{\gamma}^{(t)}) \ge 0$. 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, $\rho_t/\rho_{t-1} \ge 1 - (t+3)/(t+1)^2$. 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 $\mu_0 \ge 0$, 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.

Acknowledgments. The authors would like to thank the Editor, Associate Editor and referees for suggestions that significantly improved the paper.

Funding. The work is partially supported by the National Science Foundation.

SUPPLEMENTARY MATERIAL

Supplement to "Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning" (DOI: 10.1214/21-AOS2090SUPP; .pdf). The supplement contains technical details, algorithm outlines and computer experiments.

REFERENCES

[1] AGARWAL, A., NEGAHBAN, S. and WAINWRIGHT, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Statist.* **40** 2452–2482. MR3097609 https://doi.org/10.1214/12-AOS1032

[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. *Ann. Oper. Res.* **133** 23–46. MR2119311 https://doi.org/10.1007/s10479-004-5022-1

[3] BANERJEE, A., MERUGU, S., DHILLON, I. S. and GHOSH, J. (2005). Clustering with Bregman divergences. *J. Mach. Learn. Res.* **6** 1705–1749. MR2249870

[4] BECK, A. and TEBOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.* **2** 183–202. MR2486527 https://doi.org/10.1137/080716542

[5] BEN-TAL, A. and NEMIROVSKI, A. (2013). Optimization III: Convex analysis, nonlinear programming theory, standard nonlinear programming algorithms. Lecture Notes.

[6] BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469 https://doi.org/10.1214/08-AOS620

[7] BLUMENSATH, T. and DAVIES, M. E. (2009). Iterative hard thresholding for compressed sensing. *Appl. Comput. Harmon. Anal.* **27** 265–274. MR2559726 https://doi.org/10.1016/j.acha.2009.04.002

[8] BRÈGMAN, 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 Comput. Math. Math. Phys.* **7** 200–217. MR0215617

[9] BUNEA, F., TSYBAKOV, A. and WEGKAMP, M. (2007). Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* **1** 169–194. MR2312149 https://doi.org/10.1214/07-EJS008

[10] CANDES, E. and TAO, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313–2351. MR2382644 https://doi.org/10.1214/009053606000001523

[11] CANDES, E. J. and TAO, T. (2005). Decoding by linear programming. *IEEE Trans. Inf. Theory* **51** 4203–4215. MR2243152 https://doi.org/10.1109/TIT.2005.858979

[12] CHEN, G. and TEBOULLE, M. (1993). Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM J. Optim.* **3** 538–543. MR1230155 https://doi.org/10.1137/0803026

[13] 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.

[14] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089 https://doi.org/10.1093/biomet/81.3.425

[15] DUCHI, J., HAZAN, E. and SINGER, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.* **12** 2121–2159. MR2825422

[16] FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581 https://doi.org/10.1198/016214501753382273

[17] FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988 https://doi.org/10.1214/13-AOS1198

[18] FRANK, I. E. and FRIEDMAN, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35** 109–135.

[19] GASSO, G., RAKOTOMAMONJY, A. and CANU, S. (2009). Recovering sparse signals with a certain family of nonconvex penalties and DC programming. *IEEE Trans. Signal Process.* **57** 4686–4698. MR2722328 https://doi.org/10.1109/TSP.2009.2026004

[20] GHADIMI, S. and LAN, G. (2016). Accelerated gradient methods for nonconvex nonlinear and stochastic programming. *Math. Program.* **156** 59–99. MR3459195 https://doi.org/10.1007/s10107-015-0871-8

[21] HUBER, P. J. (1981). *Robust Statistics. Wiley Series in Probability and Mathematical Statistics*. Wiley, New York. MR0606374

[22] HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509 https://doi.org/10.1198/0003130042836

[23] HUNTER, D. R. and LI, R. (2005). Variable selection using MM algorithms. *Ann. Statist.* **33** 1617–1642. MR2166557 https://doi.org/10.1214/009053605000000200

[24] JØRGENSEN, B. (1987). Exponential dispersion models. *J. Roy. Statist. Soc. Ser. B* **49** 127–162. With discussion and a reply by the author. MR0905186

[25] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems. Lecture Notes in Math.* **2033**. Springer, Heidelberg. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d'Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. MR2829871 https://doi.org/10.1007/978-3-642-22147-7

[26] 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.

[27] LANGE, K. and ZHOU, H. (2014). MM algorithms for geometric and signomial programming. *Math. Program.* **143** 339–356. MR3152072 https://doi.org/10.1007/s10107-012-0612-1

[28] LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by nonnegative matrix factorization. *Nature* **401** 788–791.

[29] LOH, P.-L. and WAINWRIGHT, M. J. (2015). Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *J. Mach. Learn. Res.* **16** 559–616. MR3335800

[30] LOUNICI, K., PONTIL, M., VAN DE GEER, S. and TSYBAKOV, A. B. (2011). Oracle inequalities and optimal inference under group sparsity. *Ann. Statist.* **39** 2164–2204. MR2893865 https://doi.org/10.1214/11-AOS896

[31] NEMIROVSKY, A. S. and YUDIN, D. B. (1983). *Problem Complexity and Method Efficiency in Optimization*. *A Wiley-Interscience Publication*. Wiley, New York. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. MR0702836

[32] NESTEROV, Y. (2004). *Introductory Lectures on Convex Optimization*: *A Basic Course*. *Applied Optimization* **87**. Kluwer Academic, Boston, MA. MR2142598 https://doi.org/10.1007/978-1-4419-8853-9

[33] NESTEROV, YU. E. (1983). A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Sov. Math.*, *Dokl.* **27** 372–376. MR0701288

[34] NESTEROV, YU. E. (1988). An approach to constructing optimal methods for minimization of smooth convex functions. *Èkon. Mat. Metody* **24** 509–517. MR0968064

[35] PAN, W., SHEN, X. and LIU, B. (2013). Cluster analysis: Unsupervised learning via supervised learning with a non-convex penalty. *J. Mach. Learn. Res.* **14** 1865–1889. MR3104498

[36] ROCKAFELLAR, R. T. (1970). *Convex Analysis*. *Princeton Mathematical Series* **28**. Princeton Univ. Press, Princeton, NJ. MR0274683

[37] SCHMIDT, M. (2010). Graphical model structure learning with ℓ_1-regularization. Ph.D. thesis, Univ. British Columbia.

[38] SHE, Y. (2009). Thresholding-based iterative selection procedures for model selection and shrinkage. *Electron. J. Stat.* **3** 384–415. MR2501318 https://doi.org/10.1214/08-EJS348

[39] SHE, Y. (2012). An iterative algorithm for fitting nonconvex penalized generalized linear models with grouped predictors. *Comput. Statist. Data Anal.* **56** 2976–2990. MR2929353 https://doi.org/10.1016/j.csda.2011.11.013

[40] SHE, Y. (2016). On the finite-sample analysis of Θ-estimators. *Electron. J. Stat.* **10** 1874–1895. MR3522663 https://doi.org/10.1214/15-EJS1100

[41] SHE, Y., HE, Y. and WU, D. (2014). Learning topology and dynamics of large recurrent neural networks. *IEEE Trans. Signal Process.* **62** 5881–5891. MR3281530 https://doi.org/10.1109/TSP.2014.2358956

[42] SHE, Y., WANG, J., LI, H. and WU, D. (2013). Group iterative spectrum thresholding for super-resolution sparse spectral selection. *IEEE Trans. Signal Process.* **61** 6371–6386. MR3148325 https://doi.org/10.1109/TSP.2013.2281303

[43] SHE, Y., WANG, Z. and JIN, J. (2021). Supplement to "Analysis of generalized Bregman surrogate algorithms for nonsmooth nonconvex statistical learning." https://doi.org/10.1214/21-AOS2090SUPP

[44] TAO, P. D. and SOUAD, E. B. (1986). Algorithms for solving a class of nonconvex optimization problems. Methods of subgradients. In *FERMAT Days* 85: *Mathematics for Optimization* (*Toulouse*, 1985). *North-Holland Math. Stud.* **129** 249–271. North-Holland, Amsterdam. MR0874369 https://doi.org/10.1016/S0304-0208(08)72402-2

[45] TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

[46] TSENG, P. (2008). On accelerated proximal gradient methods for convex-concave optimization. Dept. Mathematics, Univ. Washington.

[47] TSYBAKOV, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[48] VAN DE GEER, S. A. and BÜHLMANN, P. (2009). On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* **3** 1360–1392. MR2576316 https://doi.org/10.1214/09-EJS506

[49] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. *Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

[50] WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

[51] WANG, Z., LIU, H. and ZHANG, T. (2014). Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.* **42** 2164–2201. MR3269977 https://doi.org/10.1214/14-AOS1238

[52] ZHANG, C., JIANG, Y. and CHAI, Y. (2010). Penalized Bregman divergence for large-dimensional regression and classification. *Biometrika* **97** 551–566. With supplementary data available online. MR2672483 https://doi.org/10.1093/biomet/asq033

[53] ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701 https://doi.org/10.1214/09-AOS729

[54] ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.* **11** 1081–1107. MR2629825

[55] ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443 https://doi.org/10.1214/009053607000000802