Gradient Descent Converges Linearly for Logistic Regression on Separable Data

Kyriakos Axiotis [1] **Maxim Sviridenko** [2]

Abstract

We show that running gradient descent with variable learning rate guarantees loss $f(\boldsymbol{x}) \leq 1.1 \cdot f(\boldsymbol{x}^*) + \varepsilon$ for the logistic regression objective, where the error ε decays exponentially with the number of iterations and polynomially with the magnitude of the entries of an arbitrary fixed solution $\boldsymbol{x}^*$. This is in contrast to the common intuition that the absence of strong convexity precludes linear convergence of first-order methods, and highlights the importance of variable learning rates for gradient descent. We also apply our ideas to sparse logistic regression, where they lead to an exponential improvement of the sparsity-error tradeoff.

1. Introduction

Logistic regression is one of the most widely used classification methods because of its simplicity, interpretability, and good practical performance. Yet, the convergence behavior of first-order methods on this task is not well understood: In practice gradient descent performs much better than what the theory predicts. In particular, a general analysis of gradient descent for smooth functions implies convergence with the error in function value decaying as $O(1/T)$. Analyses with stronger, linear convergence guarantees generally require the function to satisfy the strong convexity property, which, in contrast to other losses such as the ℓ_2 loss, the logistic loss only satisfies in a bounded set of solutions around zero. As a result, this introduces an *exponential* runtime dependency on the magnitude of the target solution (Rätsch et al., 2001; Freund et al., 2018), which is undesirable in practice. This poses a serious obstacle to obtaining high-precision solutions for logistic regression.

In fact, it was shown in (Telgarsky & Singer, 2012) that the $\mathrm{poly}(1/T)$ bound on function value convergence is tight for gradient descent on general (non-linearly separable) data. The significance of the separability of the data for convergence has also been observed in (Telgarsky, 2013; Ji & Telgarsky, 2018; Freund et al., 2018), who present convergence results based on quantitative measures of separability.

A deeper study into the structure of both the exponential and logistic losses for separable data was initiated by (Telgarsky & Singer, 2012), who showed that greedy coordinate descent achieves linear convergence with a rate that depends on the maximum linear classification margin (i.e. hard SVM margin). Unfortunately, for logistic regression, it also has a 2^m dependence on the number of examples, making it inefficient for real-world tasks. (Telgarsky, 2013) refines the results of (Telgarsky & Singer, 2012) for the exponential loss, but for logistic regression still suffers from an exponential overhead originating from the multiplicative discrepancy between the exponential and logistic losses. Interestingly, however, the authors note ((Telgarsky, 2013), Section 5) that logistic regression experiments paint a much more favorable picture than the theory predicts.

A related line of work deals with convergence to the maximum-margin classifier on linearly separable classification instances using gradient descent. (Soudry et al., 2018; Ji & Telgarsky, 2018) showed that the estimator obtained by optimizing the logistic or the exponential loss with gradient descent converges to the maximum-margin linear classifier at a rate of $O(\log\log T/\log T)$ (in ℓ_2 norm). For the exponential loss, (Nacson et al., 2019) showed that the convergence bound to the maximum margin estimator can be exponentially improved to $O(\log T/\sqrt{T})$, by using gradient descent with variable (increasing) learning rate. The authors' experiments indicate that variable step sizes could lead to a similar exponential improvements for the case of logistic regression and shallow neural networks. Recently, (Ji & Telgarsky, 2021) presented a novel primal-dual approach that proves that the latter claim indeed holds for the logistic regression and exponential objectives, obtaining a maximum-margin error decaying as $O(1/T)$, using a variable learning rate. This exponentially improved upon the results of (Soudry et al., 2018; Ji & Telgarsky, 2018).

This work was performed while the first author was at MIT. [1]Google Research, New York, NY, USA [2]Yahoo! Research, New York, NY, USA. Correspondence to: Kyriakos Axiotis <axiotis@google.com>, Maxim Sviridenko <sviri@yahooinc.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

Another approach to obtain high-precision solutions is by using second order methods, which in addition to first order (gradient) information, use second order (Hessian) information about the function. These make use of second order stability properties, such as quasi-self-concordance (Bach, 2010) combined with Newton's method (Karimireddy et al., 2018), or ball oracles (Carmon et al., 2020; Adil et al., 2021). Such approaches are generally not suitable for large-scale applications because of their reliance on repeated calls to large linear system solvers.

Our work. In this paper, we show that (under appropriate assumptions) we can get the best of both worlds of first and second order methods, thus giving a partial explanation for the excellent performance that first-order methods have for logistic regression in practice. In particular, given a binary classification instance ($\boldsymbol{A} \in \{-1,1\}^{m\times n}, \boldsymbol{b} \in \{-1,1\}^m$) with associated logistic loss $f(\boldsymbol{x}) = \sum_i \log(1+\exp(-b_i(\boldsymbol{A}\boldsymbol{x})_i))$, we show that simple variants of gradient descent return a solution with $f(\boldsymbol{x}) \le (1+\delta)\cdot f(\boldsymbol{x}^*) + \varepsilon$ after $O\left(K\left(\frac{1}{\delta} + \log\frac{f(\mathbf{0})}{\varepsilon}\right)\right)$ iterations, where $K = \mathrm{poly}(n, \|\boldsymbol{x}^*\|)$ and x^* is an arbitrary fixed solution. Even though the error still decays as $1/T$ in the worst case because of the $\frac{1}{\delta}$ dependence, the additive error is now $\delta f(\boldsymbol{x}^*)$ instead of $\delta f(\mathbf{0})$, allowing for much faster convergence when the optimal loss $f(\boldsymbol{x}^*)$ is smaller (which is our measure of linear separability of the data). For linearly separable data, i.e. as $f(\boldsymbol{x}^*)$ approaches 0, the convergence becomes linear.

Instead of properties like Lipschitzness, smoothness, strong convexity that are commonly used in the study of first order methods, we find that there are two properties that are more relevant to the structure of the logistic regression problem. The first one is *second order robustness*, which means that the Hessian is stable (in a spectral sense) in any small enough norm ball (Cohen et al., 2017). This is closely related to quasi-self-concordance, a property that has been previously used in the analysis of second order algorithms (Bach, 2010). The second property is what we call *multiplicative smoothness*, which means that the function is locally smooth, with the smoothness constant being proportional to the function value (loss). A similar property was used by (Ji & Telgarsky, 2018) to prove convergence of logistic regression to the maximum margin classifier. Together, these properties show that, as the loss decreases, the objective becomes (locally) smoother and therefore the learning rate can increase. This motivates a variable step size schedule that is inversely proportional to the loss, thus making larger steps as the solution approaches optimality. This in fact agrees with the observations of (Soudry et al., 2018; Nacson et al., 2019) on the importance of a variable learning rate. As can be seen in the toy example from (Soudry et al., 2018) in Figure 1, simply replacing the fixed learning rate η used in (Soudry et al., 2018) by an increasing learning rate $\eta \cdot f(\boldsymbol{x}^0)/f(\boldsymbol{x}^T)$ yields an exponential improvement, both in loss and distance to the maximum margin estimator.

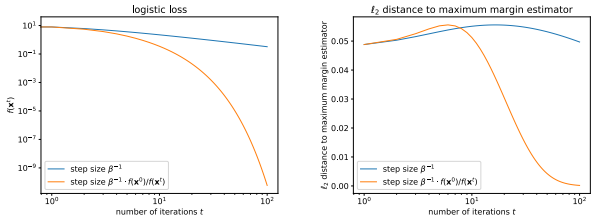

Figure 1. Comparison between fixed and increasing step sizes in the toy example from Figure 1 of (Soudry et al., 2018). The fixed step size is set to $\beta^{-1} := \|\boldsymbol{A}\|_2^{-2}$, and the increasing to $\beta^{-1} f(\boldsymbol{x}^0)/f(\boldsymbol{x}^T)$. The estimator error is defined as $\left\|\boldsymbol{x}^t/\|\boldsymbol{x}^t\|_2 - \boldsymbol{x}^*/\|\boldsymbol{x}^*\|_2\right\|_2$.

1.1. Sparse logistic regression

In practice, it is often important to force the solution of a logistic regression problem to be *sparse*, i.e. have only a few non-zero entries, which is a form of feature selection. This is because most of the features might only be marginally useful, and thus one can drastically reduce the size of the model while not significantly sacrificing the predictive performance. Apart from computational efficiency, feature selection is also important to improve interpretability and avoid overfitting.

Most progress in sparse optimization has focused on objective functions with condition number bounded by some $\kappa > 0$. Results in this line of work guarantee a solution with relaxed sparsity $s' \ge s$, where s is the target sparsity, and algorithms include lasso, orthogonal matching pursuit (OMP), and iterative hard thresholding (IHT) (Natarajan, 1995; Blumensath & Davies, 2009; Shalev-Shwartz et al., 2010; Jain et al., 2011; 2014; Axiotis & Sviridenko, 2021; 2022). The state of the art result by (Axiotis & Sviridenko, 2022) gives a sparsity of $s' = O(\kappa)\cdot s$ using a variant of the IHT algorithm.

However, the condition number of the logistic loss is unbounded, because it is not strongly convex. Therefore, these results do not directly apply, although they do apply to ℓ_2-regularized logistic regression. Some works (Van de Geer, 2008; Bunea, 2008) have analyzed lasso methods for logistic regression without condition number assumptions, and (Shalev-Shwartz et al., 2010) provides three different analyses for smooth but not strongly convex functions. These apply to logistic regression and give a sparsity of $O\left(\|\boldsymbol{x}^*\|_1^2 \frac{m}{\varepsilon}\right)$ to achieve a loss of $f(\boldsymbol{x}) \le f(\boldsymbol{x}^*) + \varepsilon$. The most practical of these is a forward greedy selection algo-

Table 1. Algorithms for logistic regression and dependence on m/ε (omitting extra $\mathrm{polylog}(m,n)$ factors). Algorithms with exponential dependences on any problem parameter are omitted. For example, the standard gradient descent analysis shows linear convergence, but with an exponential dependence on $\|\boldsymbol{x}^*\|_\infty$ (see e.g. Freund et al. (2018)).

ALGORITHM	ORDER	GUARANTEE	RUNTIME ERROR DEPENDENCE
GRADIENT DESCENT	FIRST	$f(\boldsymbol{x}) \le f(\boldsymbol{x}^*) + \varepsilon$	m/ε
ACCELERATED GRADIENT DESCENT	FIRST	$f(\boldsymbol{x}) \le f(\boldsymbol{x}^*) + \varepsilon$	$\sqrt{m/\varepsilon}$
NEWTON/TRUST REGION	SECOND	$f(\boldsymbol{x}) \le f(\boldsymbol{x}^*) + \varepsilon$	$\log(m/\varepsilon)$
THIS PAPER	FIRST	$f(\boldsymbol{x}) \le (1+\delta)\cdot f(\boldsymbol{x}^*) + \varepsilon$	$\delta^{-1} + \log(m/\varepsilon)$

Table 2. Algorithms for sparse logistic regression and asymptotic sparsity dependences.

ALGORITHM	GUARANTEE	SPARSITY	ORDER
(SHALEV-SHWARTZ ET AL., 2010)	$f(\boldsymbol{x}) \le f(\boldsymbol{x}^*) + \varepsilon$	$\|\boldsymbol{x}^*\|_1^2\, m/\varepsilon$	FIRST
THIS PAPER	$f(\boldsymbol{x}) \le (1+\delta)\cdot f(\boldsymbol{x}^*) + \varepsilon$	$\|\boldsymbol{x}^*\|_1^2\,(\delta^{-1} + \log(m/\varepsilon))$	FIRST

rithm, which is also known as greedy coordinate descent.

Our work. Using the second order stability and multiplicative smoothness properties, we show that a slight variation of greedy coordinate descent gives a sparsity of

$$O\left(\|\boldsymbol{x}^*\|_1^2\,(\delta^{-1} + \log(m/\varepsilon))\right)$$

and a loss of $f(\boldsymbol{x}) \le (1+\delta)\cdot f(\boldsymbol{x}^*) + \varepsilon$. As long as the $1+\delta$ approximation in front of $f(\boldsymbol{x}^*)$ is tolerated, as is the case when $f(\boldsymbol{x}^*) \ll m$, this implies an exponential improvement in the ε dependence from $\frac{m}{\varepsilon}$ to $\log\frac{m}{\varepsilon}$. In addition, our analysis is compatible with incorporating fully corrective steps to the algorithm. These are steps that are occasionally performed to optimize over the support of the current solution (i.e. fully optimize the weights of the currently selected features) and is often used in applications like feature selection.

2. Preliminaries

Notation. We denote $[n] = \{1, 2, \ldots, n\}$. We will use **bold** to refer to vectors or matrices. We denote by $\mathbf{0}$ the all-zero vector, $\mathbf{1}$ the all-one vector, $\boldsymbol{O}$ the all-zero matrix, and by $\boldsymbol{I}$ the identity matrix (with dimensions understood from the context). Additionally, we will denote by $\mathbf{1}_i$ the i-th basis vector, i.e. the vector that is 0 everywhere except at position i.

In order to ease notation and where not ambiguous for two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, we denote by $\boldsymbol{x}\boldsymbol{y} \in \mathbb{R}^n$ a vector with elements $(\boldsymbol{x}\boldsymbol{y})_i = x_i y_i$, i.e. the element-wise multiplication of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$. In contrast, we denote their inner product by $\langle \boldsymbol{x}, \boldsymbol{y}\rangle$ or $\boldsymbol{x}^\top \boldsymbol{y}$. Similarly, $\boldsymbol{x}^2 \in \mathbb{R}^n$ will be the element-wise square of vector $\boldsymbol{x}$. For any function $g(t)$ of a single variable, let $g(\boldsymbol{x}) \in \mathbb{R}^n$ be a vector with elements $(g(\boldsymbol{x}))_i = g(x_i)$, e.g. will use this notation when g is the sigmoid function.

For any vector $\boldsymbol{x} \in \mathbb{R}^n$ and set $S \subseteq [n]$, we denote by $\boldsymbol{x}_S$ the vector that results from $\boldsymbol{x}$ after zeroing out all the entries except those in positions given by indices in S. We will also use the notation $\nabla_S f(\boldsymbol{x}) := (\nabla f(\boldsymbol{x}))_S$ to denote the restriction of a gradient to S. We also denote by $\mathrm{supp}(\boldsymbol{x}) := \{i \in [n] \mid x_i \neq 0\}$ the *support* of $\boldsymbol{x}$.

We use the notation $\widetilde{O}(\cdot)$ to hide $\mathrm{poly}\log(n, m)$ factors in O-notation, where n is the dimension of the problem and m is the number of examples.

Norms. For any $p \in (0, \infty)$ and weight vector $\boldsymbol{w} \ge \mathbf{0}$, we define the weighted ℓ_p norm of a vector $\boldsymbol{x} \in \mathbb{R}^n$ as:

$$\|\boldsymbol{x}\|_{p,\boldsymbol{w}} = \left(\sum_i w_i x_i^p\right)^{1/p}.$$

For $p = 0$, we denote $\|\boldsymbol{x}\|_0 = |\{i \mid x_i \neq 0\}|$ to be the *sparsity* of $\boldsymbol{x}$. For $p = \infty$, we denote $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$ to be the maximum absolute value of $\boldsymbol{x}$.

For a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, we let $\|\boldsymbol{A}\|_{p\to q}$ be its p to q operator norm, defined as

$$\|\boldsymbol{A}\|_{p\to q} = \max_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_q}{\|\boldsymbol{x}\|_p}.$$

In particular, $\|\boldsymbol{A}\|_{1\to\infty}$ is equal to the largest entry of $\boldsymbol{A}$ in absolute value.

Smoothness and convexity. A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called *convex* if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have $f(\boldsymbol{y}) \ge f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x}\rangle$. Furthermore, f is called *L-Lipschitz (with respect to some norm $\|\cdot\|$)* for some real number $L > 0$ if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have

$|f(\boldsymbol{y}) - f(\boldsymbol{x})| \leq L\|\boldsymbol{y} - \boldsymbol{x}\|$, and β-*smooth* if for any $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have $\|\nabla f(\boldsymbol{y}) - f(\boldsymbol{x})\|_* \leq \beta\|\boldsymbol{y} - \boldsymbol{x}\|^2$, where $\|\cdot\|_*$ is the *dual norm* of $\|\cdot\|$. If f is only β-smooth along s-sparse directions (i.e. only for $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ such that $\|\boldsymbol{y} - \boldsymbol{x}\|_0 \leq s$), then we call f β-smooth *at sparsity level* s and denote the smallest such β by β_s and call it the *restricted smoothness constant* (at sparsity level s).

3. Logistic Regression Analysis via Multiplicative Smoothness

In the logistic regression problem, our goal is to minimize the function $f(\boldsymbol{x}) = \sum_{i=1}^{m} \log(1 + e^{-(\boldsymbol{A}\boldsymbol{x})_i})$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is a data matrix[1].

Our starting point, as is usually the case with first-order methods, will be the second order Taylor expansion of f:

$$f(\boldsymbol{x} + \widetilde{\boldsymbol{x}}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}}\rangle + \frac{1}{2}\langle \widetilde{\boldsymbol{x}}, \nabla^2 f(\bar{\boldsymbol{x}})\widetilde{\boldsymbol{x}}\rangle\,, \tag{1}$$

where, by the mean value theorem for twice continuously differentiable functions, $\bar{\boldsymbol{x}}$ is entry-wise between $\boldsymbol{x}$ and $\boldsymbol{x}' = \boldsymbol{x} + \widetilde{\boldsymbol{x}}$, and $\nabla^2 f(\bar{\boldsymbol{x}})$ is the Hessian of f at $\bar{\boldsymbol{x}}$.

Second-order robustness. In fact, as long as the step $\widetilde{\boldsymbol{x}}$ is not too large, the Hessian at $\bar{\boldsymbol{x}}$ will not differ much (spectrally) from the Hessian at $\boldsymbol{x}$. This is because of the following property of the logistic function called *second order robustness* (Cohen et al., 2017), which is also very closely related to quasi-self-concordance (Bach, 2010).

Definition 3.1 (Second-order robustness)**.** *A twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called* q-second order robust *with respect to a norm $\|\cdot\|$ if its Hessian is stable in any $(1/q)$-sized $\|\cdot\|$-ball, i.e. for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ such that $\|\boldsymbol{x}' - \boldsymbol{x}\| \leq 1/q$, we have $\frac{1}{2}\nabla^2 f(\boldsymbol{x}) \preceq \nabla^2 f(\boldsymbol{x}') \preceq 2\nabla^2 f(\boldsymbol{x})$.*

It is not hard to see that f is $2M$-second order robust with respect to the ℓ_1 norm, where M is an upper bound on the entries of $\boldsymbol{A}$ in absolute value.

Lemma 3.2 (Second-order robustness of the logistic loss)**.** *The function $f(\boldsymbol{x}) = \sum_{i=1}^{m} \log(1 + e^{-(\boldsymbol{A}\boldsymbol{x})_i})$ is $2M$-second order robust with respect to the ℓ_1 norm, where M is the largest entry of $\boldsymbol{A}$ in absolute value.*

Proof. For any $\boldsymbol{x} \in \mathbb{R}^n$, we have $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^\top \mathrm{diag}(w(\boldsymbol{A}\boldsymbol{x}))\boldsymbol{A}$, where $w(\boldsymbol{A}\boldsymbol{x}) = \sigma(\boldsymbol{A}\boldsymbol{x})(1-\sigma(\boldsymbol{A}\boldsymbol{x}))$ and $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function, that we extend to vectors by applying it elementwise. We define $r(t) := \log w(t)$, which is a 1-Lipschitz function, because

$$r'(t) = \frac{w'(t)}{w(t)} = \frac{w(t)\cdot(1 - 2\sigma(t))}{w(t)} = 1 - 2\sigma(t)\,,$$

whose absolute value is always bounded by 1. Therefore, for any $|t' - t| \leq 1/2$, we have

$$\begin{aligned} &|r(t') - r(t)| \leq 1/2 \\ \Leftrightarrow &\left|\log\frac{w(t')}{w(t)}\right| \leq 1/2 \\ \Rightarrow &\frac{1}{2}w(t) \leq w(t') \leq 2w(t)\,, \end{aligned}$$

and consequently for any $\boldsymbol{x}, \boldsymbol{x}'$ such that

$$\|\boldsymbol{x}' - \boldsymbol{x}\|_1 \leq \frac{1}{2M} \Rightarrow \|\boldsymbol{A}\boldsymbol{x}' - \boldsymbol{A}\boldsymbol{x}\|_\infty \leq 1/2\,,$$

we have that $\frac{1}{2}w(\boldsymbol{A}\boldsymbol{x}) \leq w(\boldsymbol{A}\boldsymbol{x}') \leq 2w(\boldsymbol{A}\boldsymbol{x})$. This immediately implies that $\frac{1}{2}\nabla^2 f(\boldsymbol{x}) \preceq \nabla^2 f(\boldsymbol{x}') \preceq 2\nabla^2 f(\boldsymbol{x})$, concluding the proof. □

Because of Lemma 3.2, (1) implies the much simpler

$$f(\boldsymbol{x} + \widetilde{\boldsymbol{x}}) \leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}}\rangle + \langle \widetilde{\boldsymbol{x}}, \nabla^2 f(\boldsymbol{x})\widetilde{\boldsymbol{x}}\rangle\,, \tag{2}$$

as long as $\|\widetilde{\boldsymbol{x}}\|_1 \leq 1/(2M)$.

Multiplicative smoothness. We can easily calculate that $\nabla f(\boldsymbol{x}) = -\boldsymbol{A}^\top(\boldsymbol{1} - \sigma(\boldsymbol{A}\boldsymbol{x}))$, where $\sigma(t) = 1/(1 + e^{-t})$ is the sigmoid function, and $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^\top \mathrm{diag}(w(\boldsymbol{A}\boldsymbol{x}))\boldsymbol{A}$, where $w(\boldsymbol{A}\boldsymbol{x}) = \sigma(\boldsymbol{A}\boldsymbol{x})(\boldsymbol{1}-\sigma(\boldsymbol{A}\boldsymbol{x}))$ are diagonal weights. Now, we should note that the second order term of (2) can be re-written as $\langle w(\boldsymbol{A}\boldsymbol{x}), (\boldsymbol{A}\widetilde{\boldsymbol{x}})^2\rangle$. This term, whose magnitude is what will determine the step size of the algorithm and in turn the bound on the total number of iterations, becomes smaller as the weights $w(\boldsymbol{A}\boldsymbol{x})$ become smaller. The crucial observation is that these weights are bounded in a way that depends on the logistic loss of the solution $\boldsymbol{x}$, as shown in the lemma below:

Lemma 3.3 (Sum of second derivatives of the logistic function)**.** *Let $f(\boldsymbol{x}) = \sum_{i=1}^{m} \log(1 + e^{-(\boldsymbol{A}\boldsymbol{x})_i})$ and $w(t) = \sigma(t)(1 - \sigma(t))$, where σ is the sigmoid function. Then,*

$$\sum_{i=1}^{m} w((\boldsymbol{A}\boldsymbol{x})_i) \leq f(\boldsymbol{x})\,. \tag{3}$$

Proof. We have $w(t) = \sigma(t)(1 - \sigma(t)) \leq 1 - \sigma(t) \leq -\log\sigma(t) = \log(1 + e^{-t})$, where we used the inequality $\log p \leq p - 1$ for all $p > 0$. □

[1] This formulation is without loss of generality, because we can incorporate the binary ± 1 labels into the matrix $\boldsymbol{A}$ and assume that all the labels are positive.

In other words, as the loss decreases, f becomes *smoother* (in an appropriate sense). This is what allows the algorithm to employ a step size that is *inversely proportional* to the loss. A similar observation has been made in (Ji & Telgarsky, 2018). The above discussion motivates the following definition of *multiplicative smoothness*. This is related to the usual definition of smoothness but also incorporates the property that the function becomes smoother as the loss decreases.

Definition 3.4 (Multiplicative smoothness)**.** *We call a differentiable function $f : \mathbb{R}^n \to \mathbb{R}_{>0}$ μ-*multiplicatively smooth *with respect to a norm $\|\cdot\|$, if for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^n$ we have*

$$\frac{\widetilde{\boldsymbol{x}}^\top \nabla^2 f(\boldsymbol{x}) \widetilde{\boldsymbol{x}}}{f(\boldsymbol{x})} \leq \mu \|\widetilde{\boldsymbol{x}}\|^2 \ .$$

Our use of a general norm is not an over-generalization, since as we will see the ℓ_1 norm is more suitable for sparse logistic regression, and the ℓ_2 norm is more suitable for the unrestricted case. In fact, it can be proved that the logistic loss is M^2-multiplicatively smooth with respect to the ℓ_1 norm, where we remind that M is a bound on the entries of $\boldsymbol{A}$ in absolute value:

Lemma 3.5 (Multiplicative smoothness of the logistic loss)**.** *The function $f(\boldsymbol{x}) = \sum_{i=1}^{m} \log(1 + e^{-(\boldsymbol{A}\boldsymbol{x})_i})$ is M^2-multiplicatively smooth with respect to the ℓ_1 norm, where M is the largest entry of $\boldsymbol{A}$ in absolute value.*

Proof. For any $\boldsymbol{x}, \widetilde{\boldsymbol{x}} \in \mathbb{R}^n$, using the fact that $\nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}^\top \operatorname{diag}(w(\boldsymbol{A}\boldsymbol{x}))\, \boldsymbol{A}$, where w is as defined in Lemma 3.3, we have

$$\begin{aligned}
\widetilde{\boldsymbol{x}}^\top \nabla^2 f(\boldsymbol{x}) \widetilde{\boldsymbol{x}} &= \sum_{i=1}^{m} w(\boldsymbol{A}\boldsymbol{x})_i (\boldsymbol{A}\widetilde{\boldsymbol{x}})_i^2 \\
&\leq \sum_{i=1}^{m} w(\boldsymbol{A}\boldsymbol{x})_i \|\boldsymbol{A}\widetilde{\boldsymbol{x}}\|_\infty^2 \\
&\leq M^2 \sum_{i=1}^{m} w(\boldsymbol{A}\boldsymbol{x})_i \|\widetilde{\boldsymbol{x}}\|_1^2 \\
&\leq M^2 f(\boldsymbol{x}) \|\widetilde{\boldsymbol{x}}\|_1^2 \ ,
\end{aligned}$$

where the second to last inequality follows from the fact that the entries of $\boldsymbol{A}$ are bounded by M, and the last inequality from Lemma 3.3. □

In the following sections, we will see how the second order robustness and multiplicative smoothness properties play into the design and analysis of algorithms for sparse and general logistic regression.

4. Sparse logistic regression

As we saw, the logistic loss is $2M$-second order robust and M^2-multiplicatively smooth with respect to the ℓ_1 norm. This is an ideal norm for *sparse* logistic regression, where in addition to minimizing the loss we want to restrict the solution to have few non-zero entries. In particular, it yields a variant of the ℓ_1 gradient descent algorithm (aka greedy coordinate descent), which is presented in Algorithm 1.

Algorithm 1 Greedy Coordinate Descent

function GreedyCoordinateDescent($\boldsymbol{x}^0, T, M, B$)
  Let $f(\boldsymbol{x}) := \sum_{i=1}^{m} \log(1 + e^{-b_i(\boldsymbol{A}\boldsymbol{x})_i})$
  for $t = 0, \ldots, T-1$ **do**
    For all $i \in [n]$ define $\zeta_i = \begin{cases} \lambda_t & \text{if } x_i^t = 0 \\ 0 & \text{if } |x_i^t| \geq B \text{ and } \nabla_i f(\boldsymbol{x}^t) \cdot x_i^t < 0 \\ 1 & \text{otherwise} \end{cases}$
    $i' \leftarrow \operatorname{argmax}_i \{\zeta_i |\nabla_i f(\boldsymbol{x}^t)|\}$
    $\eta \leftarrow \left(2M^2 f(\boldsymbol{x}^t)\right)^{-1}$
    $x_{i'}^{t+1} \leftarrow x_{i'}^t - \eta \nabla_{i'} f(\boldsymbol{x}^t)$
  end for
  return $\boldsymbol{x}^T$
end function

The first thing that should be noted about this algorithm is the crucial parameters λ_t. These parameters offer a quantitative threshold between sparsity and speed of convergence. In particular, when λ_t is 1, then all entries (regardless of whether they are zero or not) are treated the same. When $\lambda_t \ll 1$, on the other hand, the gradient entries corresponding to zero entries are discounted by a factor $\ll 1$, thus making the algorithm less eager to update these as opposed to non-zero entries, whose update doesn't increase sparsity.

We are ready for the main theorem of this section. In the proof, which can be found in Appendix A.2.1, we present an analysis of Algorithm 1 for sparse logistic regression. In addition to an upper bound $B \geq \|\boldsymbol{x}^*\|_\infty$, it also requires an approximation B_1 of $\|\boldsymbol{x}^*\|_1$. One possible approach is to approximate it by B, but in practice this would be a learning rate hyperparameter to be tuned. We note $\boldsymbol{x}^*$ is an arbitrary fixed solution (not necessarily optimal), whose entries are bounded by B in absolute value.

Theorem 4.1 (Sparse logistic regression)**.** *Given a binary classification instance ($\boldsymbol{A} \in [-M, M]^{m \times n}, \boldsymbol{b} \in \{1, -1\}^m$) and for any solution $\boldsymbol{x}^* \in [-B, B]^n$ with $M \geq \max\{\|\boldsymbol{x}^*\|_1^{-1}, B^{-1}\}$ [2] and a known parameter $B_1 \in \left[\frac{1}{C}\|\boldsymbol{x}^*\|_1, \|\boldsymbol{x}^*\|_1\right]$ for some $C \geq 1$, Algorithm 1 with $\lambda_t = \min\{B_1 / \|\boldsymbol{x}^t\|_1, 1\}$, initial solution $\boldsymbol{x}^0 \in \mathbb{R}^n$,*

[2] The theorem can be stated without this additional constraint, but we include it because it makes the bounds considerably simpler.

and error tolerance $0 < \varepsilon < m/2$ *returns a solution* $\boldsymbol{x}$ *with*

$$f(\boldsymbol{x}) \leq (1+\delta) \cdot f(\boldsymbol{x}^*) + \varepsilon$$

and sparsity

$$\begin{aligned} s' &:= \|\boldsymbol{x}\|_0 \\ &= O\left(\|\boldsymbol{x}^*\|_1^2 M^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right) \end{aligned}$$

in

$$\begin{aligned} T = O\Big(&\left(\|\boldsymbol{x}\|_0^2 + \|\boldsymbol{x}^*\|_0^2\right) \\ &M^2 B^2 C^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\Big) \end{aligned}$$

iterations, for any choice of $\delta \in (0,1)$. *Each iteration consists of evaluating the logistic regression gradient* ∇f *plus* $O(m+n)$ *additional time.*

If the parameters M, B, C are bounded by $\widetilde{O}(1)$, we get a cleaner statement.

Corollary 4.2. *If* $M, B, C \leq \widetilde{O}(1)$ *and* $\boldsymbol{x}^*$ *is* s*-sparse, then Algorithm 1 with* $\lambda_t = \min\{B_1 / \|\boldsymbol{x}^t\|_1, 1\}$ *returns a solution* $\boldsymbol{x}$ *with*

$$f(\boldsymbol{x}) \leq 1.1 \cdot f(\boldsymbol{x}^*) + \varepsilon$$

and sparsity

$$\begin{aligned} s' &:= \|\boldsymbol{x}\|_0 \\ &= \widetilde{O}\left(s^2 \log \frac{1}{\varepsilon}\right) \end{aligned}$$

in

$$T = \widetilde{O}\left(s^4 \log^3 \frac{1}{\varepsilon}\right)$$

iterations.

Corollary 4.2 follows from Theorem 4.1 by applying the following series of inequalities:

$$\begin{aligned} \|\boldsymbol{x}\|_0^2 &\leq \widetilde{O}\left(\|\boldsymbol{x}^*\|_1^4 \log^2 \frac{1}{\varepsilon}\right) \\ &\leq \widetilde{O}\left(B^4 \|\boldsymbol{x}^*\|_0^4 \log^2 \frac{1}{\varepsilon}\right) \\ &\leq \widetilde{O}\left(\|\boldsymbol{x}^*\|_0^4 \log^2 \frac{1}{\varepsilon}\right). \end{aligned}$$

It is useful to compare these results to the results of (Shalev-Shwartz et al., 2010) for sparse optimization of general smooth convex functions. Even though those results achieve the stronger error bound of $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*) + \varepsilon$, the sparsity of the final solution is in the order of $s^2 \frac{m}{\varepsilon}$, which has an exponentially worse error dependence than $s^2 \log \frac{m}{\varepsilon}$. Therefore, if the approximation rate $(1+\delta)$ is tolerable in front of $f(\boldsymbol{x}^*)$, then one can obtain exponentially faster sparsity and convergence.

If we are willing to perform fully corrective steps as described in Algorithm 2, then we can get a cleaner and slightly simpler analysis without dependence on parameters B, B_1, C. This is presented in Theorem 4.3 and proved in Appendix A.2.2. Fully corrective steps can be useful when there is an efficient (dense) optimization algorithm and one wishes to use it as a black box for sparse optimization. In practice, one does not need to perform a full correction, but only a small number of corrective gradient steps over the current support of the solution.

Algorithm 2 Greedy coordinate descent with fully corrective steps

function FullyCorrectiveGreedyCD($\boldsymbol{x}^0, T$)
  Let $f(\boldsymbol{x}) := \sum_{i=1}^{m} \log(1 + e^{-b_i(\boldsymbol{A}\boldsymbol{x})_i})$
  $S^0 \leftarrow \text{supp}(\boldsymbol{x}^0)$
  for $t = 0, \ldots, T-1$ **do**
    $i' \leftarrow \text{argmax}_i \{|\nabla_i f(\boldsymbol{x}^t)|\}$
    $S^{t+1} \leftarrow S^t \cup \{i'\}$
    $\boldsymbol{x}^{t+1} \leftarrow \underset{\boldsymbol{x}:\text{supp}(\boldsymbol{x}) \subseteq S^{t+1}}{\text{argmin}} f(\boldsymbol{x})$
  end for
  return $\boldsymbol{x}^T$
end function

Theorem 4.3 (Sparse logistic regression with fully corrective steps)**.** *Given a binary classification instance* $(\boldsymbol{A} \in [-M, M]^{m \times n}, \boldsymbol{b} \in \{1, -1\}^m)$ *and for any solution* $\boldsymbol{x}^* \in \mathbb{R}^n$*, Algorithm 2 with error tolerance* $0 < \varepsilon < m/2$ *and initial solution* $\boldsymbol{x}^0$ *returns a solution* $\boldsymbol{x}$ *with*

$$f(\boldsymbol{x}) \leq (1+\delta) \cdot f(\boldsymbol{x}^*) + \varepsilon$$

and sparsity

$$\begin{aligned} s' &:= \|\boldsymbol{x}\|_0 \\ &= \|\boldsymbol{x}^0\|_0 + O\left(\|\boldsymbol{x}^*\|_1^2 M^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right) \end{aligned}$$

in $T = O\left(\|\boldsymbol{x}^*\|_1^2 M^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right)$ *iterations, for any choice of* $\delta \in (0,1)$. *Each iteration consists of evaluating the logistic regression gradient* ∇f*, solving a logistic regression problem on at most* s' *variables, plus* $O(m+n)$ *additional time.*

5. Dense logistic regression

In this section, our goal is to minimize the logistic function f without any constraint on the sparsity of the solution. The

results of Section 4 applied to a full sparsity of n already imply Corollary 5.1.

Corollary 5.1 (Dense logistic regression)**.** *Given a binary classification instance* ($\boldsymbol{A} \in [-M, M]^{m\times n}, \boldsymbol{b} \in \{-1, 1\}^m$) *and for any solution* $\boldsymbol{x}^* \in [-B, B]^n$ *with* $M \geq \max\left\{\|\boldsymbol{x}^*\|_1^{-1}, B^{-1}\right\}$*, Algorithm 1 with* $\lambda_t = 1$ *for all* t*, initial solution* $\boldsymbol{x}^0 \in \mathbb{R}^n$*, and error tolerance* $0 < \varepsilon < m/2$ *returns a solution* $\boldsymbol{x}$ *with*

$$f(\boldsymbol{x}) \leq (1+\delta)\cdot f(\boldsymbol{x}^*) + \varepsilon$$

in

$$T = O\left(n^2M^2B^2\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right).$$

iterations, for any choice of $\delta \in (0, 1)$*. Additionally,* $\|\boldsymbol{x}\|_\infty \leq B + \frac{1}{2M}$*. Each iteration consists of evaluating the logistic regression gradient* ∇f *plus* $O(m+n)$ *additional time.*

Interestingly, even if we don't impose a sparsity constraint, the worst case analysis in Corollary 5.1 still obtains its bounds by using a greedy coordinate descent step, as in Section 4. This is because it is a manifestation of ℓ_1 gradient descent (aka steepest descent), which is rooted in the fact that the multiplicative smoothness of f is with respect to the ℓ_1 norm. Greedy coordinate descent is favorable for sparse optimization, because it only updates one coordinate at a time. On the other hand, based on practical intuitions, we would expect (ℓ_2-based) gradient descent to perform better than greedy coordinate descent if no sparsity constraint is imposed. This is because greedy coordinate descent only updates one coordinate at a time, while gradient descent uses the full gradient.

In the rest of this section, we attempt to bridge this mismatch between worst case analysis and practice. We show that, under appropriate assumptions, gradient descent can be proved to converge significantly faster than greedy coordinate descent. We leave providing a theoretical grounding for these assumptions as an interesting open question for future research. We now outline the core ideas.

ℓ_2 multiplicative smoothness. First, we can verify that the logistic loss does have the multiplicative smoothness condition with respect to the ℓ_2 norm, albeit in an almost trivial sense:

$$\begin{aligned}\langle \boldsymbol{w}(\boldsymbol{x}), (\boldsymbol{A}\widetilde{\boldsymbol{x}})^2\rangle &\leq \|\boldsymbol{w}(\boldsymbol{x})\|_1 \|\boldsymbol{A}\widetilde{\boldsymbol{x}}\|_\infty^2 \\ &\leq f(\boldsymbol{x}) \|\boldsymbol{A}\|_{2\to\infty}^2 \|\widetilde{\boldsymbol{x}}\|_2^2 \\ &\leq f(\boldsymbol{x})\beta \|\widetilde{\boldsymbol{x}}\|_2^2 .\end{aligned}$$

Here, using the inequality $\|\boldsymbol{A}\|_{2\to\infty}^2 \leq \|\boldsymbol{A}\|_2^2 := \beta$ implies β-multiplicative smoothness with respect to the ℓ_2 norm.

Table 3. Upper bounds on the quantity $\left\langle \boldsymbol{w}(\boldsymbol{x}), (\boldsymbol{A}\nabla f(\boldsymbol{x}))^2\right\rangle / \left(f(\boldsymbol{x})m^{-1}\|\boldsymbol{A}\nabla f(\boldsymbol{x})\|_2^2\right)$. Shown here is the maximum of this over $\boldsymbol{x}$ being one of the first 1000 iterates starting from $\boldsymbol{x}^0 = \boldsymbol{0}$.

Dataset	Max ratio
letter	0.40
rcv1.test	0.36
ijcnn1	0.47
vehv2binary	0.37
magic04	0.37
skin	0.44
w8all	0.40
shuttle.binary	0.37
kddcup04.phy	0.36
kddcup04.bio	0.48
census	0.50
adult	0.40
poker	0.36
nomao	0.50
covtype	0.36

Unfortunately, this is not significantly better than the ℓ_1 case: The number of iterations will be proportional to $\beta\|\boldsymbol{x}^*\|_2^2$, which can be $\gg m$. As it turns out, however, many real logistic regression instances exhibit the ℓ_2 multiplicative smoothness property with significantly better constants. In our experiments we found that along the path of gradients encountered by gradient descent in a variety of instances, the following property was true:

$$\left\langle \boldsymbol{w}(\boldsymbol{x}), (\boldsymbol{A}\nabla f(\boldsymbol{x}))^2\right\rangle \leq f(\boldsymbol{x})\beta m^{-1}\|\nabla f(\boldsymbol{x})\|_2^2$$

This is an *effective* βm^{-1}-multiplicative smoothness property, because it is only assumed to be true for $\boldsymbol{x}$'s encountered by the gradient descent algorithm. As such, it is an empirical property. In order to check our hypothesis, we have run the gradient descent algorithm with the step sizes that are implied by Theorem 5.2, which we will see later. For each of the 15 experiments, we have run gradient descent for 1000 iterations, and calculated the maximum of the following quantity, over all iterations:

$$\frac{\left\langle \boldsymbol{w}(\boldsymbol{x}), (\boldsymbol{A}\nabla f(\boldsymbol{x}))^2\right\rangle}{f(\boldsymbol{x})m^{-1}\|\boldsymbol{A}\nabla f(\boldsymbol{x})\|_2^2}.$$

If this is bounded by 1, and using the fact that $\|\boldsymbol{A}\nabla f(\boldsymbol{x})\|_2^2 \leq \beta\|\nabla f(\boldsymbol{x})\|_2^2$, this implies that f is effectively βm^{-1}-multiplicatively smooth with respect to the ℓ_2 norm. Indeed, as we can see in Table 3, these values are indeed less than 1 for all datasets and all iterations.

In the following, our plan is to prove convergence, *assuming* that f has the multiplicative smoothness property with the

constants in our hypothesis above. Under this assumption, we can now prove a much stronger convergence theorem (here we are also using the fact that $M^2 \leq \beta$ to replace $2M$-by $2\sqrt{\beta}$-second order robustness):

Theorem 5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}_{>0}$ be a convex function that is γ-second order robust and μ-multiplicatively smooth with respect to the ℓ_2 norm. Let $\boldsymbol{x}^0 \in \mathbb{R}^n$ be an initial solution and $\boldsymbol{x}^* \in \mathbb{R}^n$ be an arbitrary solution, where $R := \left\|\boldsymbol{x}^0 - \boldsymbol{x}^*\right\|_2$. Then, gradient descent with step size $\eta_t = \min\left\{\frac{1}{2\mu f(\boldsymbol{x}^t)}, \frac{1}{\gamma\|\nabla f(\boldsymbol{x}^t)\|_2}\right\}$ returns a solution with*

$$f(\boldsymbol{x}) \leq (1+\delta)f(\boldsymbol{x}^*) + \varepsilon$$

after

$$T = O\Big(\mu R^2 \left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right) + \gamma R \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\Big)$$

iterations. If $\gamma \leq 2\sqrt{\beta}$ and $\mu \leq \beta m^{-1}$ for some $\beta > 0$, then the number of iterations becomes

$$T = O\Big(\frac{\beta R^2}{m} \left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right) + \sqrt{\beta} R \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\Big)$$

Theorem 5.2 is proved in Appendix B.2.

6. Numerical Example

In order to numerically validate our algorithm, we run logistic regression on the UCI adult binary classification dataset. In order to simulate a separable dataset, we first run gradient descent on the whole data, and then discard the misclassified data points. This gives us a separable dataset. Then, we run two variants of gradient descent: One with constant step size given by β^{-1}, and one with increasing step size given by $\eta_t = \beta^{-1} f(\boldsymbol{x}^0)/f(\boldsymbol{x}^t)$, with no other change. This is motivated by our findings, which suggest that the step size should increase proportionally to the decrease of the loss. As we can see in Figure 2, the error in the case of fixed step size decays as $\mathrm{poly}(1/t)$, while in the case of increasing step size we have linear convergence (albeit with a low rate because the margins are in the order of 10^{-6}).

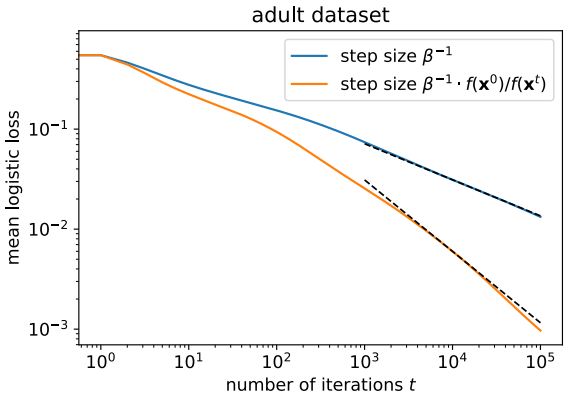

Figure 2. Comparison of fixed vs increasing step size on logistic regression on adult dataset

7. Acknowledgments

We would like to thank the anonymous reviewers for valuable feedback, and pointing out a simplification to the choice of step size in Algorithm 1.

References

Adil, D., Bullins, B., and Sachdeva, S. Unifying width-reduced methods for quasi-self-concordant optimization. *Advances in Neural Information Processing Systems*, 34, 2021.

Axiotis, K. and Sviridenko, M. Sparse convex optimization via adaptively regularized hard thresholding. *Journal of Machine Learning Research*, 22:1–47, 2021.

Axiotis, K. and Sviridenko, M. Iterative hard thresholding with adaptive regularization: Sparser solutions without sacrificing runtime. *arXiv preprint arXiv:2204.08274*, 2022.

Bach, F. Self-concordant analysis for logistic regression. *Electronic Journal of Statistics*, 4:384–414, 2010.

Blumensath, T. and Davies, M. E. Iterative hard thresholding for compressed sensing. *Applied and computational harmonic analysis*, 27(3):265–274, 2009.

Bunea, F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, 2:1153–1194, 2008.

Carmon, Y., Jambulapati, A., Jiang, Q., Jin, Y., Lee, Y. T., Sidford, A., and Tian, K. Acceleration with a ball optimization oracle. *Advances in Neural Information Processing Systems*, 33:19052–19063, 2020.

Cohen, M. B., Madry, A., Tsipras, D., and Vladu, A. Matrix scaling and balancing via box constrained newton's method and interior point methods. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 902–913. IEEE, 2017.

Freund, R. M., Grigas, P., and Mazumder, R. Condition number analysis of logistic regression, and its implications for standard first-order solution methods. *arXiv preprint arXiv:1810.08727*, 2018.

Jain, P., Tewari, A., and Dhillon, I. S. Orthogonal matching pursuit with replacement. In *Advances in neural information processing systems*, pp. 1215–1223, 2011.

Jain, P., Tewari, A., and Kar, P. On iterative hard thresholding methods for high-dimensional m-estimation. In *Advances in Neural Information Processing Systems*, pp. 685–693, 2014.

Ji, Z. and Telgarsky, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300*, 2018.

Ji, Z. and Telgarsky, M. Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pp. 772–804. PMLR, 2021.

Karimireddy, S. P., Stich, S. U., and Jaggi, M. Global linear convergence of newton's method without strong-convexity or lipschitz gradients. *arXiv preprint arXiv:1806.00413*, 2018.

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 3420–3428. PMLR, 2019.

Natarajan, B. K. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

Rätsch, G., Mika, S., and Warmuth, M. K. On the convergence of leveraging. *Advances in Neural Information Processing Systems*, 14, 2001.

Shalev-Shwartz, S., Srebro, N., and Zhang, T. Trading accuracy for sparsity in optimization problems with sparsity constraints. *SIAM Journal on Optimization*, 20(6): 2807–2832, 2010.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

Telgarsky, M. Margins, shrinkage, and boosting. In *International Conference on Machine Learning*, pp. 307–315. PMLR, 2013.

Telgarsky, M. and Singer, Y. A primal-dual convergence analysis of boosting. *Journal of Machine Learning Research*, 13(3), 2012.

Van de Geer, S. A. High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2): 614–645, 2008.

A. Missing Proofs from Section 4

A.1. Main lemma on coordinate updates

Lemma A.1 (Gradient lower bound)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable convex function and let $\boldsymbol{x} \in [-B', B']^n$, $\boldsymbol{x}^* \in [-B, B]^n$ be two solutions for some parameters $B' \geq B > 0$. For all $i \in [n]$ we define*

$$\zeta_i = \begin{cases} \lambda & \text{if } x_i = 0 \\ 0 & \text{if } |x_i| \geq B \text{ and } \nabla_i f(\boldsymbol{x}) \cdot x_i < 0 \\ 1 & \text{otherwise} \end{cases}$$

where $0 < \lambda \leq 1$, and let $i^* = \operatorname{argmax}_i \{\zeta_i |\nabla_i f(\boldsymbol{x})|\}$. Then, at least one of the following is true:

- $|\nabla_{i^*} f(\boldsymbol{x})| \geq \dfrac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^*\|_1 + \lambda \|\boldsymbol{x}\|_1}$
- $|\nabla_{i^*} f(\boldsymbol{x})| \geq \dfrac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{\lambda^{-1} \|\boldsymbol{x}^*\|_1 + \|\boldsymbol{x}\|_1}$ *and* $x_{i^*} \neq 0$.

Proof. Let $S = \{i \mid x_i \neq 0\}$ and $F = \{i \mid |x_i| < B \text{ or } \nabla_i f(\boldsymbol{x}) \cdot x_i \geq 0\}$. By convexity of f, we have

$$\begin{aligned} f(\boldsymbol{x}^*) &\geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x} \rangle \\ &\geq f(\boldsymbol{x}) + \langle \nabla_F f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x} \rangle \\ &= f(\boldsymbol{x}) + \langle \nabla_F f(\boldsymbol{x}), \boldsymbol{x}^* \rangle - \langle \nabla_{S \cap F} f(\boldsymbol{x}), \boldsymbol{x} \rangle \\ &\geq f(\boldsymbol{x}) - \|\nabla_F f(\boldsymbol{x})\|_\infty \|\boldsymbol{x}^*\|_1 - \|\nabla_{S \cap F} f(\boldsymbol{x})\|_\infty \|\boldsymbol{x}\|_1 \ , \end{aligned}$$

where the first inequality holds because for any $i \in [n] \backslash F$, by the fact that $|x_i^*| \leq B$ and the definition of F,

$$\nabla_i f(\boldsymbol{x}) \left(\boldsymbol{x}^* - \boldsymbol{x}\right)_i \geq |\nabla_i f(\boldsymbol{x})| \left(-B + B\right) = 0 \, .$$

Therefore

$$\|\nabla_F f(\boldsymbol{x})\|_\infty \|\boldsymbol{x}^*\|_1 + \|\nabla_{S \cap F} f(\boldsymbol{x})\|_\infty \|\boldsymbol{x}\|_1 \geq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \, . \tag{4}$$

Now, if $i^* \notin S$, which also implies $x_{i^*} = 0$, by definition of the ζ_i's and i^* we have

$$\lambda \|\nabla_F f(\boldsymbol{x})\|_\infty = \lambda |\nabla_{i^*} f(\boldsymbol{x})| \geq \|\nabla_{S \cap F} f(\boldsymbol{x})\|_\infty \ .$$

and so (4) implies

$$\begin{aligned} &|\nabla_{i^*} f(\boldsymbol{x})| \|\boldsymbol{x}^*\|_1 + \lambda |\nabla_{i^*} f(\boldsymbol{x})| \|\boldsymbol{x}\|_1 \geq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \\ \Rightarrow &|\nabla_{i^*} f(\boldsymbol{x})| \geq \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^*\|_1 + \lambda \|\boldsymbol{x}\|_1} \, . \end{aligned}$$

Otherwise if $i^* \in S$, we have

$$\|\nabla_{S \cap F} f(\boldsymbol{x})\|_\infty = |\nabla_{i^*} f(\boldsymbol{x})| \geq \lambda \|\nabla_F f(\boldsymbol{x})\|_\infty \ ,$$

and so (4) implies

$$\begin{aligned} &\lambda^{-1} |\nabla_{i^*} f(\boldsymbol{x})| \|\boldsymbol{x}^*\|_1 + |\nabla_{i^*} f(\boldsymbol{x})| \|\boldsymbol{x}\|_1 \geq f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \\ \Rightarrow &|\nabla_{i^*} f(\boldsymbol{x})| \geq \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{\lambda^{-1} \|\boldsymbol{x}^*\|_1 + \|\boldsymbol{x}\|_1} \, . \end{aligned}$$

□

Lemma A.2 (Coordinate update)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}_{>0}$ be a twice continuously differentiable convex function that is 2γ-second order robust and γ^2-multiplicatively smooth with respect to the ℓ_1 norm, for some $\gamma > 0$. Let $\boldsymbol{x} \in [-B', B']^n$ be a suboptimal solution such that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}^*)$, where $\boldsymbol{x}^* \in [-B, B]^n$ is some unknown solution with $\gamma \|\boldsymbol{x}^*\|_1 \geq 1$, and $B' \geq B > 0$ are some parameters. We make the update*

$$\boldsymbol{x}' = \boldsymbol{x} - \eta \nabla_i f(\boldsymbol{x}) \mathbf{1}_i \,,$$

where i is picked as in Lemma A.1 for some parameter $\lambda \in (0, 1)$ and $\eta = 0.5 \min\left\{\frac{1}{\gamma^2 f(\boldsymbol{x})}, \frac{1}{\gamma |\nabla_i f(\boldsymbol{x})|}\right\}$ is a step size. Then, at least one of the following is true about the progress in decreasing f:

- $f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \dfrac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\gamma^2 f(\boldsymbol{x}) \left(\|\boldsymbol{x}^*\|_1 + \lambda \|\boldsymbol{x}\|_1\right)^2}$
- $f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \dfrac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\gamma^2 f(\boldsymbol{x}) \left(\lambda^{-1} \|\boldsymbol{x}^*\|_1 + \|\boldsymbol{x}\|_1\right)^2}$ *and* $x_i \neq 0 \,,$

and the norm of the new solution is bounded as $\|\boldsymbol{x}'\|_\infty \leq \max\{B', B + \frac{1}{2\gamma}\}$. In the case that $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$ we have $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$.

Proof. We first consider a generic update $\boldsymbol{x}' = \boldsymbol{x} + \widetilde{\boldsymbol{x}}$. By Taylor's theorem and the fact that f is twice continuously differentiable, we have

$$f(\boldsymbol{x}') = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} \rangle + \frac{1}{2} \langle \widetilde{\boldsymbol{x}}, \nabla^2 f(\bar{\boldsymbol{x}}) \widetilde{\boldsymbol{x}} \rangle \,,$$

for some $\bar{\boldsymbol{x}}$ that is entrywise between $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Since f is 2γ-second-order-robust and γ^2-multiplicatively-smooth with respect to the ℓ_1 norm, as long as the update is bounded in ℓ_1 norm as

$$\|\widetilde{\boldsymbol{x}}\|_1 \leq 1/(2\gamma) \tag{5}$$

we have

$$\begin{aligned} f(\boldsymbol{x}') &\leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} \rangle + \langle \widetilde{\boldsymbol{x}}, \nabla^2 f(\boldsymbol{x}) \widetilde{\boldsymbol{x}} \rangle \\ &\leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} \rangle + \gamma^2 f(\boldsymbol{x}) \|\widetilde{\boldsymbol{x}}\|_1^2 \,. \end{aligned}$$

Note that the right hand side is minimized for

$$\widetilde{\boldsymbol{x}} = -\frac{H_1(\nabla f(\boldsymbol{x}))}{2\gamma^2 f(\boldsymbol{x})} \,,$$

where H_1 is the hard thresholding operator that zeroes out all but the top entry in absolute value. This is a coordinate descent step. Our step will be slightly more careful so that it doesn't unnecessarily increase the sparsity of $\boldsymbol{x}$. We consider the following coordinate step

$$\widetilde{\boldsymbol{x}} = -\eta \nabla_i f(\boldsymbol{x}) \mathbf{1}_i \,,$$

where $\eta > 0$ and i are as defined in the lemma statement. We now have a function value decrease of

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \left(\eta - \eta^2 \gamma^2 f(\boldsymbol{x})\right) (\nabla_i f(\boldsymbol{x}))^2 \,.$$

The term $\eta - \eta^2 \gamma^2 f(\boldsymbol{x})$ is maximized at $\eta = \frac{1}{2\gamma^2 f(\boldsymbol{x})}$. In addition, to stay in the ℓ_1 neighborhood where the Hessian is stable, we need to satisfy (5) by making sure that $\eta \leq \frac{1}{2\gamma |\nabla_i f(\boldsymbol{x})|}$. Based on these requirements, we pick

$$\eta = \min\left\{\frac{1}{2\gamma^2 f(\boldsymbol{x})}, \frac{1}{2\gamma |\nabla_i f(\boldsymbol{x})|}\right\}$$

and conclude that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \min\left\{\frac{1}{4\gamma^2 f(\boldsymbol{x})}, \frac{1}{4\gamma|\nabla_i f(\boldsymbol{x})|}\right\}(\nabla_i f(\boldsymbol{x}))^2$$
$$= \min\left\{\frac{(\nabla_i f(\boldsymbol{x}))^2}{4\gamma^2 f(\boldsymbol{x})}, \frac{|\nabla_i f(\boldsymbol{x})|}{4\gamma}\right\}.$$

Note that this is always ≥ 0 and so we have $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$ even if $f(\boldsymbol{x}) < f(\boldsymbol{x}^*)$. We now take two cases and use the two bullets of Lemma A.1 accordingly.

Case 1: $x_i = 0$**.** The first bullet of Lemma A.1 has to be true, i.e.

$$|\nabla_i f(\boldsymbol{x})| \geq \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{\|\boldsymbol{x}^*\|_1 + \lambda \|\boldsymbol{x}\|_1}.$$

Therefore,

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \min\left\{\frac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\gamma^2 f(\boldsymbol{x})\left(\|\boldsymbol{x}^*\|_1 + \lambda\|\boldsymbol{x}\|_1\right)^2}, \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{4\gamma\left(\|\boldsymbol{x}^*\|_1 + \lambda\|\boldsymbol{x}\|_1\right)}\right\}$$
$$= \frac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\gamma^2 f(\boldsymbol{x})\left(\|\boldsymbol{x}^*\|_1 + \lambda\|\boldsymbol{x}\|_1\right)^2},$$

where we used the facts that $f(\boldsymbol{x}) - f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ and $\gamma\|\boldsymbol{x}^*\|_1 \geq 1$.

Case 2: $x_i \neq 0$**.** If the first bullet of Lemma A.1 is true, we can proceed as in the previous case. Otherwise, we use the second bullet of Lemma A.1 and similarly get

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \frac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\gamma^2 f(\boldsymbol{x})\left(\lambda^{-1}\|\boldsymbol{x}^*\|_1 + \|\boldsymbol{x}\|_1\right)^2}.$$

Finally, in order to bound $\|\boldsymbol{x}'\|_\infty$, we first note that $\|\boldsymbol{x}\|_\infty \leq B'$. Now, by our choice of i we have that either $|x_i| < B$, or $\nabla_i f(\boldsymbol{x}) \cdot x_i > 0$. In the first case, we have

$$|x_i'| \leq |x_i| + |\widetilde{x}_i| < B + \frac{1}{2\gamma},$$

where we used (5). Otherwise, we have that $|x_i| \geq B$ and $\nabla_i f(\boldsymbol{x}) \cdot x_i > 0$. This implies that x_i and $\widetilde{x}_i$ have different signs, so

$$|x_i'| = |x_i + \widetilde{x}_i| \leq \max\{|x_i|, |\widetilde{x}_i|\} \leq \max\left\{B', \frac{1}{2\gamma}\right\}.$$

Therefore, in any case we have $|x_i'| \leq \max\left\{B', B + \frac{1}{2\gamma}\right\}$. □

A.2. Theorems

A.2.1. Proof of Theorem 4.1

Proof. We will apply Lemma A.2 for T iterations to obtain solutions $\boldsymbol{x}^0, \ldots, \boldsymbol{x}^T$, for some T that will be defined later. The step size parameter $\lambda_t < 1$ disincentivizes updating zero entries of the solution vector. The logistic function f is $2M$-second order robust and M^2-multiplicatively smooth with respect to the ℓ_1 norm (Lemmas 3.2 and 3.5), so Lemma A.2 can be applied with $\gamma = M$ and $B' = B + \frac{1}{2M}$.

Also, we can simplify the step size used in Lemma A.2 to $\frac{1}{2M^2 f(\boldsymbol{x})}$. This is because

$$\|\nabla f(\boldsymbol{x})\|_\infty = \left\|\boldsymbol{A}^\top(1 - \sigma(\boldsymbol{A}\boldsymbol{x}))\right\|_\infty$$
$$\leq M\|1 - \sigma(\boldsymbol{A}\boldsymbol{x})\|_1$$
$$\leq M f(\boldsymbol{x}),$$

where we used the fact that $1 - \sigma(t) = 1/(1+e^t) \leq \log(1+e^{-t})$. The inequality holds because the function $g(t) = (1+e^t)\log(1+e^{-t})$ is decreasing ($g'(t) = e^t \log(1+e^{-t}) - 1 \leq 0$) and goes to 1 as $t \to \infty$.

The progress bound of Lemma A.2 depends on $\|\boldsymbol{x}^t\|_1$. Based on the guarantee of Lemma A.2 about $\|\boldsymbol{x}^t\|_\infty$, we can derive the following bound:

$$\begin{aligned}\left\|\boldsymbol{x}^t\right\|_1 &\leq \left\|\boldsymbol{x}^t\right\|_0 \left\|\boldsymbol{x}^t\right\|_\infty \\ &\leq \left\|\boldsymbol{x}^t\right\|_0 \left(B + \frac{1}{2M}\right) \\ &\leq \left\|\boldsymbol{x}^t\right\|_0 (3/2)B\,.\end{aligned}$$

We can now bound the sparsity. Note that the sparsity increases by at most 1 every time the first bullet of Lemma A.2 is true, and does not increase when the second bullet is true. Therefore, the progress in each sparsity-increasing iteration, based on our choice of λ_t, is

$$\begin{aligned}f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) &\geq \frac{\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right)^2}{4f(\boldsymbol{x}^t)M^2\left(\|\boldsymbol{x}^*\|_1 + \lambda_t \|\boldsymbol{x}^t\|_1\right)^2} \\ &\geq \frac{\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right)^2}{4f(\boldsymbol{x}^t)M^2\left(\|\boldsymbol{x}^*\|_1 + B_1\right)^2} \\ &\geq \frac{\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right)^2}{16f(\boldsymbol{x}^t)M^2\|\boldsymbol{x}^*\|_1^2}\,.\end{aligned}$$

The following auxiliary lemma will help us turn this into a convergence result:

Lemma A.3. *Consider a function $f : \mathbb{R}^n \to \mathbb{R}_{>0}$ and a sequence $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots$ of iterates such that for all t*

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) \geq \alpha \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{f(\boldsymbol{x}^t)} \tag{6}$$

for some $\boldsymbol{x}^*$ with $f(\boldsymbol{x}^*) \leq \min_t f(\boldsymbol{x}^t)$. Then,

$$f(\boldsymbol{x}^T) \leq (1+\delta)f(\boldsymbol{x}^*) + \varepsilon$$

after

$$T \leq 2\alpha^{-1}\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)$$

iterations.

Proof. Let $\bar{t}$ be the smallest $t \geq 0$ for which $f(\boldsymbol{x}^{\bar{t}}) \leq 2f(\boldsymbol{x}^*)$ or $f(\boldsymbol{x}^{\bar{t}}) \leq f(\boldsymbol{x}^*) + \varepsilon$, and let $\bar{t} = \infty$ if this never happens. Therefore, for all $t < \bar{t}$ we have $f(\boldsymbol{x}^t) \geq 2f(\boldsymbol{x}^*) \Rightarrow \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^t)} \geq \frac{1}{2}$, Then, (6) implies

$$\begin{aligned}f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) &\geq \frac{\alpha}{2}\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right) \\ \Rightarrow f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*) &\leq \left(1 - \frac{\alpha}{2}\right)\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right)\,.\end{aligned}$$

Applying these for all t, we get

$$f(\boldsymbol{x}^{\bar{t}}) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{\alpha}{2}\right)^{\bar{t}}\left(f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)\right)\,,$$

which directly implies that

$$\bar{t} \leq 2\alpha^{-1}\log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\,.$$

Now, let $t \geq \bar{t}$. If $f(\boldsymbol{x}^{\bar{t}}) \leq f(\boldsymbol{x}^*) + \varepsilon$ we are done and there is nothing to prove. Otherwise, $f(\boldsymbol{x}^t) \leq f(\boldsymbol{x}^{\bar{t}}) \leq 2f(\boldsymbol{x}^*)$ and therefore (6) implies

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) \geq \frac{\alpha}{2f(\boldsymbol{x}^*)}(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2 .$$

This is a well known recurrence that leads to the bound

$$f(\boldsymbol{x}^T) \leq f(\boldsymbol{x}^*) + \frac{2f(\boldsymbol{x}^*)}{\alpha(T-\bar{t})} = \left(1 + \frac{2}{\alpha(T-\bar{t})}\right) f(\boldsymbol{x}^*) ,$$

therefore

$$T \leq \frac{2}{\alpha\delta} + \bar{t} \leq 2\alpha^{-1}\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right) .$$

□

Lemma A.3 implies that the total number of sparsity-increasing iterations is

$$s' := \left\|\boldsymbol{x}^T\right\|_0 \leq 32\left\|\boldsymbol{x}^*\right\|_1^2 M^2\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right) .$$

Now it remains to bound the total number of iterations in which the second bullet of Lemma A.2 is true. These iterations do not increase the sparsity. We have

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) \geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{4f(\boldsymbol{x}^t)M^2\left(\lambda_t^{-1}\left\|\boldsymbol{x}^*\right\|_1 + \left\|\boldsymbol{x}^t\right\|_1\right)^2} .$$

Note that

$$\begin{aligned}\lambda_t^{-1}\left\|\boldsymbol{x}^*\right\|_1 &= \max\left\{\left\|\boldsymbol{x}^t\right\|_1\left\|\boldsymbol{x}^*\right\|_1/B_1, \left\|\boldsymbol{x}^*\right\|_1\right\} \\ &\leq \max\left\{C\left\|\boldsymbol{x}^t\right\|_1, \left\|\boldsymbol{x}^*\right\|_1\right\} ,\end{aligned}$$

If $C\left\|\boldsymbol{x}^t\right\|_1 \geq \left\|\boldsymbol{x}^*\right\|_1$, then

$$\begin{aligned}f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) &\geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{4f(\boldsymbol{x}^t)M^2(C+1)^2\left\|\boldsymbol{x}^t\right\|_1^2} \\ &\geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{9f(\boldsymbol{x}^t)M^2B^2(C+1)^2\left\|\boldsymbol{x}^t\right\|_0^2} \\ &\geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{9f(\boldsymbol{x}^t)M^2B^2(C+1)^2(s')^2} .\end{aligned}$$

By Lemma A.3, there can only be

$$\begin{aligned}T &= 9M^2B^2(C+1)^2(s')^2\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right) \\ &= O\left((s')^2M^2B^2C^2\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right)\end{aligned}$$

iterations. Similarly if $C\left\|\boldsymbol{x}^t\right\|_1 < \left\|\boldsymbol{x}^*\right\|_1$ we have at most

$$O\left(\left((s')^2 + \left\|\boldsymbol{x}^*\right\|_0^2\right)M^2B^2\left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right)$$

such iterations, so the result follows.

□

A.2.2. PROOF OF THEOREM 4.3

Proof. We move similarly to the proof of Theorem 4.1, but now we can strengthen Lemma A.2 because $\boldsymbol{x}^t$ is fully corrected for all t, i.e. $\nabla_i f(\boldsymbol{x}^t) = 0$ for all $i \in \operatorname{supp}(\boldsymbol{x}^t)$. As in the proof of Lemma A.2, we can lower bound the amount of progress as a function of $\|\nabla f(\boldsymbol{x}^t)\|_\infty$ as follows:

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) \geq \frac{(\nabla_i f(\boldsymbol{x}^t))^2}{4M^2 f(\boldsymbol{x}^t)}\,.$$

Now, by convexity of f we have

$$\langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^t - \boldsymbol{x}^* \rangle \geq f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\,. \tag{7}$$

Because of fully corrective steps we have $\langle \nabla f(\boldsymbol{x}^t), \boldsymbol{x}^t \rangle = 0$, and so the left hand side of (7) is upper bounded by $\|\nabla f(\boldsymbol{x}^t)\|_\infty \|\boldsymbol{x}^*\|_1$. As a result, we have

$$\left\|\nabla f(\boldsymbol{x}^t)\right\|_\infty^2 \geq \frac{\left(f(\boldsymbol{x}) - f(\boldsymbol{x}^*)\right)^2}{\|\boldsymbol{x}^*\|_1^2}\,,$$

and so we get the progress bound of

$$f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) \geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{4M^2 f(\boldsymbol{x}^t)\,\|\boldsymbol{x}^*\|_1^2}\,.$$

By Lemma A.3 (similarly to the proof of Theorem 4.1), this progress bound leads to a sparsity of

$$s' := \left\|\boldsymbol{x}^T\right\|_0 \leq O\left(\|\boldsymbol{x}^*\|_1^2 M^2 \left(\frac{1}{\delta} + \log\frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right)$$

and the same number of iterations. □

B. Missing Proofs from Section 5

B.0.1. PROOF OF COROLLARY 5.1

Proof. We will apply Lemma A.2 for T iterations to obtain solutions $\boldsymbol{x}^0, \ldots, \boldsymbol{x}^T$, where for some T that will be defined later. The logistic function f is $2M$-second order robust and M^2-multiplicatively smooth with respect to the ℓ_1 norm, so Lemma A.2 can be applied with $\gamma = M$ and $B' = B + \frac{1}{2M}$.

We get the following bound on the ℓ_1 norm of $\boldsymbol{x}^t$ at all times:

$$\left\|\boldsymbol{x}^t\right\|_1 \leq n\left\|\boldsymbol{x}^t\right\|_\infty \leq n\left(B + \frac{1}{2M}\right) \leq (3/2)nB\,.$$

Let $\bar{t}$ be the smallest $t \geq 0$ for which $f(\boldsymbol{x}^{\bar{t}}) \leq 2f(\boldsymbol{x}^*)$ or $f(\boldsymbol{x}^{\bar{t}}) \leq f(\boldsymbol{x}^*) + \varepsilon$, and let $\bar{t} = \infty$ if this never happens. Therefore, for all $t < \bar{t}$ we have $f(\boldsymbol{x}^t) \geq 2f(\boldsymbol{x}^*) \Rightarrow \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{f(\boldsymbol{x}^t)} \geq \frac{1}{2}$, and so the statement of Lemma A.2 gives:

$$\begin{aligned} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) &\geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{8M^2(\|\boldsymbol{x}^*\|_1 + \|\boldsymbol{x}^t\|_1)^2} \\ &\geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{8n^2M^2(B + (3/2)B)^2} \\ &\geq \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{50n^2M^2B^2}\,, \end{aligned}$$

where we used the fact that $\|\boldsymbol{x}^*\|_1 \leq n\|\boldsymbol{x}^*\|_\infty \leq nB$. Equivalently,

$$f(\boldsymbol{x}^{t+1}) - f(\boldsymbol{x}^*) \leq \left(1 - \frac{1}{50n^2M^2B^2}\right)\left(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)\right),$$

and summing up these for $t \in \{0, 1, \ldots, \bar{t} - 1\}$, we get

$$\begin{aligned} f(\boldsymbol{x}^{\bar{t}}) - f(\boldsymbol{x}^*) &\leq \left(1 - \frac{1}{50n^2M^2B^2}\right)^{\bar{t}} (f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)) \\ &\leq \varepsilon \,, \end{aligned}$$

as long as

$$\bar{t} \geq 50n^2M^2B^2 \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon} \,,$$

therefore we conclude that $\bar{t}$ is at most this quantity.

Now we consider the iterations $t \geq \bar{t}$. If $f(\boldsymbol{x}^{\bar{t}}) \leq f(\boldsymbol{x}^*) + \varepsilon$ there are no such iterations and we are done. Therefore we have that $f(\boldsymbol{x}^{\bar{t}}) \leq 2f(\boldsymbol{x}^*)$. We again use Lemma A.2 for all such t, which gives

$$\begin{aligned} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) &\geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{4M^2 f(\boldsymbol{x}^t)(\|\boldsymbol{x}^*\|_1 + \|\boldsymbol{x}^t\|_1)^2} \\ &\geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{25 f(\boldsymbol{x}^t) n^2 M^2 B^2} \\ &\geq \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{50 f(\boldsymbol{x}^*) n^2 M^2 B^2} \,. \end{aligned}$$

By known convergence results, this recurrence leads to the bound

$$\begin{aligned} f(\boldsymbol{x}^T) &\leq f(\boldsymbol{x}^*) + \frac{100 f(\boldsymbol{x}^*) n^2 M^2 B^2}{T - \bar{t}} \\ &= f(\boldsymbol{x}^*) \left(1 + \frac{100 n^2 M^2 B^2}{T - \bar{t}}\right) \,, \end{aligned}$$

implying that $f(\boldsymbol{x}^T) \leq (1 + \delta) f(\boldsymbol{x}^*)$ after

$$T - \bar{t} = O\left(n^2 M^2 B^2 \frac{1}{\delta}\right)$$

additional iterations after $\bar{t}$. Therefore, the total number of iterations to achieve $f(\boldsymbol{x}^T) \leq (1 + \delta) \cdot f(\boldsymbol{x}^*) + \varepsilon$ is

$$O\left(n^2 M^2 B^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon}\right)\right) \,.$$

□

B.1. Gradient update lemma

Lemma B.1. *Let $f : \mathbb{R}^n \to \mathbb{R}_{>0}$ be a twice continuously differentiable convex function that is γ-second order robust and μ-multiplicatively smooth with respect to a norm $\|\cdot\|$ for some $\gamma, \mu > 0$. Given a solution $\boldsymbol{x} \in \mathbb{R}^n$, we make the update*

$$\boldsymbol{x}' = \boldsymbol{x} - \eta \boldsymbol{g} \,,$$

where

$$\boldsymbol{g} = \operatorname{argmin}_{\boldsymbol{g} \in \mathbb{R}^n} \langle \nabla f(\boldsymbol{x}), -\boldsymbol{g} \rangle + \frac{1}{2} \|\boldsymbol{g}\|^2 \tag{8}$$

and $\eta \leq \min\left\{\frac{1}{2\mu f(\boldsymbol{x})}, \frac{1}{\gamma \|\nabla f(\boldsymbol{x})\|}\right\}$ is a step size. Then, the progress in decreasing f is:

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \frac{\eta}{2} \|\boldsymbol{g}\|^2 \,.$$

Proof. We first consider a generic update $\boldsymbol{x}' = \boldsymbol{x} + \widetilde{\boldsymbol{x}}$. By Taylor's theorem and the fact that f is twice continuously differentiable, we have

$$f(\boldsymbol{x}') = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} \rangle + \frac{1}{2} \langle \widetilde{\boldsymbol{x}}, \nabla^2 f(\bar{\boldsymbol{x}}) \widetilde{\boldsymbol{x}} \rangle \,,$$

for some $\bar{\boldsymbol{x}}$ that is entrywise between $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Since f is γ-second order robust and μ-multiplicatively-smooth with respect to the norm $\|\cdot\|$, as long as the update is bounded as

$$\|\widetilde{\boldsymbol{x}}\| \leq 1/\gamma \,, \tag{9}$$

we have

$$\begin{aligned} f(\boldsymbol{x}') &\leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} \rangle + \langle \widetilde{\boldsymbol{x}}, \nabla^2 f(\boldsymbol{x}) \widetilde{\boldsymbol{x}} \rangle \\ &\leq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \widetilde{\boldsymbol{x}} \rangle + \mu f(\boldsymbol{x}) \|\widetilde{\boldsymbol{x}}\|^2 \,. \end{aligned}$$

Note that the right hand side is minimized for

$$\widetilde{\boldsymbol{x}} = -\frac{1}{2\mu f(\boldsymbol{x})} \boldsymbol{g} \,.$$

In addition, to stay in the neighborhood where the Hessian is stable, we need to satisfy (9). Based on these requirements, we make the update $\widetilde{\boldsymbol{x}} = -\eta \boldsymbol{g}$, where

$$\eta = \min \left\{ \frac{1}{2\mu f(\boldsymbol{x})}, \frac{1}{\gamma \|\boldsymbol{g}\|} \right\} .$$

Also, by the first-order optimality condition of (8) it directly follows that $\langle \nabla f(\boldsymbol{x}), \boldsymbol{g} \rangle = \|\boldsymbol{g}\|^2$. Therefore,

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{x}') &\geq \eta \langle \nabla f(\boldsymbol{x}), \boldsymbol{g} \rangle - \eta^2 \mu f(\boldsymbol{x}) \|\boldsymbol{g}\|^2 \\ &= \eta \left(1 - \eta \mu f(\boldsymbol{x})\right) \|\boldsymbol{g}\|^2 \\ &\geq \frac{\eta}{2} \|\boldsymbol{g}\|^2 \,, \end{aligned}$$

where the last inequality follows by our choice of step size. □

Lemma B.2 (Gradient update)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}_{>0}$ be a twice continuously differentiable convex function that is γ-second order robust and μ-multiplicatively smooth with respect to the ℓ_2 norm for some $\gamma, \mu > 0$. Let $\boldsymbol{x} \in \mathbb{R}^n$ be a solution such that $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$, where $\boldsymbol{x}^* \in \mathbb{R}^n$ is an unknown solution with $\|\boldsymbol{x} - \boldsymbol{x}^*\|_2 \leq R$ for some $R > 0$. We make the update*

$$\boldsymbol{x}' = \boldsymbol{x} - \eta \nabla f(\boldsymbol{x}) \,,$$

where $\eta = \min \left\{ \frac{1}{2\mu f(\boldsymbol{x})}, \frac{1}{\gamma \|\nabla f(\boldsymbol{x})\|_2} \right\}$ is a step size. Then, the progress in decreasing f is:

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \min \left\{ \frac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\mu f(\boldsymbol{x}) R^2}, \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{2\gamma R} \right\} .$$

Additionally, as long as $\boldsymbol{x}'$ is still suboptimal with respect to $\boldsymbol{x}^*$, i.e. $f(\boldsymbol{x}') > f(\boldsymbol{x}^*)$, the distance to $\boldsymbol{x}^*$ decreases: $\|\boldsymbol{x}' - \boldsymbol{x}^*\|_2 \leq \|\boldsymbol{x} - \boldsymbol{x}^*\|_2$. Finally, if $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$, then $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$.

Proof. We apply Lemma B.1 with the ℓ_2 norm, which gives $\boldsymbol{g} = \nabla f(\boldsymbol{x})$ and so

$$\begin{aligned} f(\boldsymbol{x}) - f(\boldsymbol{x}') &\geq \frac{\eta}{2} \|\nabla f(\boldsymbol{x})\|_2^2 \\ &= \min \left\{ \frac{1}{4\mu f(\boldsymbol{x})}, \frac{1}{2\gamma \|\nabla f(\boldsymbol{x})\|_2} \right\} \|\nabla f(\boldsymbol{x})\|_2^2 \\ &= \min \left\{ \frac{\|\nabla f(\boldsymbol{x})\|_2^2}{4\mu f(\boldsymbol{x})}, \frac{\|\nabla f(\boldsymbol{x})\|_2}{2\gamma} \right\} . \end{aligned}$$

This takes care of the case $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$, since it shows that $f(\boldsymbol{x}') \leq f(\boldsymbol{x})$. Now we deal with the case $f(\boldsymbol{x}) > f(\boldsymbol{x}^*)$. By convexity we have

$$\begin{aligned} f(\boldsymbol{x}^*) &\geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{x}^* - \boldsymbol{x} \rangle \\ &\geq f(\boldsymbol{x}) - \|\nabla f(\boldsymbol{x})\|_2 \|\boldsymbol{x}^* - \boldsymbol{x}\|_2 \\ &\geq f(\boldsymbol{x}) - \|\nabla f(\boldsymbol{x})\|_2 R \,, \end{aligned}$$

which gives

$$\|\nabla f(\boldsymbol{x})\|_2^2 \geq \frac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{R^2} \,,$$

and so

$$f(\boldsymbol{x}) - f(\boldsymbol{x}') \geq \min \left\{ \frac{(f(\boldsymbol{x}) - f(\boldsymbol{x}^*))^2}{4\mu f(\boldsymbol{x}) R^2}, \frac{f(\boldsymbol{x}) - f(\boldsymbol{x}^*)}{2\gamma R} \right\} .$$

For the norm bound, we suppose that $f(\boldsymbol{x}') > f(\boldsymbol{x}^*)$ (otherwise we are done). We have

$$\begin{aligned} &\|\boldsymbol{x}' - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \\ &= \|\boldsymbol{x}' - \boldsymbol{x}\|_2^2 + 2\langle \boldsymbol{x} - \boldsymbol{x}^*, \boldsymbol{x}' - \boldsymbol{x} \rangle \\ &= \eta^2 \|\nabla f(\boldsymbol{x})\|_2^2 - 2\eta \langle \boldsymbol{x} - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}) \rangle \,. \end{aligned}$$

Now, note that

$$\frac{\eta}{2} \|\nabla f(\boldsymbol{x})\|_2^2 \leq f(\boldsymbol{x}) - f(\boldsymbol{x}') \leq f(\boldsymbol{x}) - f(\boldsymbol{x}^*)$$

and by convexity $\langle \boldsymbol{x} - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}) \rangle \geq f(\boldsymbol{x}) - f(\boldsymbol{x}^*)$, so

$$\begin{aligned} &\|\boldsymbol{x}' - \boldsymbol{x}^*\|_2^2 - \|\boldsymbol{x} - \boldsymbol{x}^*\|_2^2 \\ &= \eta^2 \|\nabla f(\boldsymbol{x})\|_2^2 - 2\eta \langle \boldsymbol{x} - \boldsymbol{x}^*, \nabla f(\boldsymbol{x}) \rangle \\ &\leq 0 \,. \end{aligned}$$

□

B.2. Proof of Theorem 5.2

Proof. We repeatedly use Lemma B.2 to obtain iterates $\boldsymbol{x}^0, \boldsymbol{x}^1, \ldots, \boldsymbol{x}^T$. Note that as long as $f(\boldsymbol{x}^t) > f(\boldsymbol{x}^*)$, we have $\|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2 \leq \left\|\boldsymbol{x}^0 - \boldsymbol{x}^*\right\|_2 := R$. We have

$$\begin{aligned} f(\boldsymbol{x}^t) - f(\boldsymbol{x}^{t+1}) &\geq \min \left\{ \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{4\mu f(\boldsymbol{x}^t) \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2^2}, \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{2\gamma \|\boldsymbol{x}^t - \boldsymbol{x}^*\|_2} \right\} \\ &\geq \min \left\{ \frac{(f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*))^2}{4\mu f(\boldsymbol{x}^t) R^2}, \frac{f(\boldsymbol{x}^t) - f(\boldsymbol{x}^*)}{2\gamma R} \right\} . \end{aligned}$$

Let there be T_1 iterations where the first branch of the minimum is smaller, and T_2 where the second one is smaller, and suppose that $f(\boldsymbol{x}^T) > (1+\delta) f(\boldsymbol{x}^*) + \varepsilon$, where $T = T_1 + T_2$. By Lemma A.3 we immediately obtain that

$$T_1 \leq 8\mu R^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon} \right) .$$

For T_2, a standard recurrence shows that

$$T_2 \leq 4\gamma R \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon} \,.$$

Therefore the total number of iterations is bounded by

$$T \leq O\left(\mu R^2 \left(\frac{1}{\delta} + \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon} \right) + \gamma R \log \frac{f(\boldsymbol{x}^0) - f(\boldsymbol{x}^*)}{\varepsilon} \right) .$$

Replacing $\mu \leq \beta m^{-1}$ and $\gamma \leq 2\sqrt{\beta}$ we obtain the desired result. □