Condition Number Analysis of Logistic Regression, and its Implications for Standard First-Order Solution Methods

Robert M. Freund[*] Paul Grigas[†] Rahul Mazumder[‡]

October 23, 2018

Abstract

Logistic regression is one of the most popular methods in binary classification, wherein estimation of model parameters is carried out by solving the maximum likelihood (ML) optimization problem, and the ML estimator is defined to be the optimal solution of this problem. It is well known that the ML estimator exists when the data is non-separable, but fails to exist when the data is linearly separable. First-order methods are the algorithms of choice for solving large-scale instances of the logistic regression problem. In this paper, we introduce a pair of condition numbers that measure the degree of non-separability or separability of a given dataset in the setting of binary classification, and we study how these condition numbers relate to and inform the properties and the convergence guarantees of first-order methods. When the training data is non-separable, we show that the degree of non-separability naturally enters the analysis and informs the properties and convergence guarantees of two standard first-order methods: steepest descent (for any given norm) and stochastic gradient descent. Expanding on the work of Bach, we also show how the degree of non-separability enters into the analysis of linear convergence of steepest descent (without needing strong convexity), as well as the adaptive convergence of stochastic gradient descent. When the training data is separable, first-order methods rather curiously have good empirical success – a behavior that is not well understood in theory. In the case of separable data, we demonstrate how the degree of separability enters into the analysis of ℓ_2 steepest descent and stochastic gradient descent for delivering approximate-maximum-margin solutions with associated computational guarantees as well. This suggests that first-order methods can lead to statistically meaningful solutions in the separable case, even though the ML solution does not exist.

[*] MIT Sloan School of Management, 77 Massachusetts Avenue, Cambridge, MA 02139 (mailto: rfreund@mit.edu). This author's research is supported by AFOSR Grant No. FA9550-15-1-0276 and the MIT-Chile-Pontificia Universidad Católica de Chile Seed Fund, and the MIT-Belgium Université Catholique de Louvain Fund.

[†] Department of Industrial Engineering and Operations Research, University of California, Berkeley, CA 94720 (mailto: pgrigas@berkeley.edu). This author's research is supported by NSF Awards CCF-1755705 and CMMI-1762744.

[‡] MIT Sloan School of Management, 77 Massachusetts Avenue, Cambridge, MA 02139 (mailto: rahulmaz@mit.edu). This author's research was partially supported by onr-n000141512342, onr-n000141812298 (YIP) and NSF-IIS-1718258.

1 Introduction

Logistic regression is arguably one of the most popular methods for binary classification – in contrast to SVM-based classifiers, logistic regression provides estimates of the probability of class membership, which is useful for uncertainty quantification and statistical inference. Moreover, the logistic loss function (and its multiclass extension, the cross-entropy loss) is an essential ingredient of several popular and powerful statistical methods, such as boosting [11], kernel methods [32], and deep learning [13].

Let us recall the setting of binary classification and logistic regression. Given a binary response $y \in \{-1, 1\}$ and feature vector $\mathbf{x} \in \mathbb{R}^p$, we consider a probability model of the form:

$$\mathbb{P}\left(y = +1 \mid \mathbf{x}\right) \ = \ \frac{1}{1 + \exp(-\beta^T \mathbf{x})} \ , \tag{1}$$

for a vector of coefficients $\beta \in \mathbb{R}^p$. The standard procedure for estimating the (unknown) coefficients β in (1) based on a given dataset of n observations $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ is to apply the principle of maximum likelihood (ML) estimation. After some basic algebraic manipulations, maximum likelihood estimation yields the following convex optimization problem:

$$\begin{aligned} \text{LR} \ : \quad L_n^* := \ \min_{\beta} \quad & L_n(\beta) \ := \ \tfrac{1}{n} \textstyle\sum_{i=1}^n \ln\left(1 + \exp\left(-y_i \beta^T \mathbf{x}_i\right)\right) \\ \text{s.t.} \quad & \beta \in \mathbb{R}^p \ . \end{aligned} \tag{2}$$

The above objective function $L_n(\cdot)$ is referred to as the logistic loss function, the i^{th} term of which measures the value of the logistic loss $t \mapsto \ln(1 + \exp(-t))$ on the i^{th} observation $(\mathbf{x}_i, y_i)$.

Due in part to classical studies [1, 30] pertaining to the existence of a ML estimator for the logistic regression problem as well as the prevalence of support vector machines (SVMs), it has become natural and customary to characterize binary classification problems in terms of their *separability* properties. More formally, a given dataset is either *separable*, in which case the set of observations with $y_i = +1$ may be separated from the set of observations with $y_i = -1$ by a (linear) hyperplane, or is *non-separable*, in which case no such linear separator exists. Earlier work in the statistics literature by [1, 30] have shown that a ML estimator for logistic regression exists when the data is non-separable, and it does not exist when the data is separable. Fairly recently, [8] studies phase transitions of the existence of a solution when the features arise from a Gaussian ensemble. A related important theme pertains to algorithms for computing a solution to problem LR. Informally, it is well-known that computational schemes for fitting a logistic regression model by solving the problem LR are "well-behaved" when the dataset is non-separable. Indeed, by simply examining (1), it is evident that if the data is truly generated according to a probabilistic model satisfying (1), then with enough samples the dataset will eventually be non-separable; therefore non-separability is somehow an essential characteristic of logistic regression. On the other hand, separability of a dataset (especially when $n > p$) suggests that there actually is a linear model that can discriminate between the two classes with high accuracy. In this case, it is well understood that the SVM method may be used to identify a "good" linear separator. Furthermore, although the logistic loss function encourages models that linearly separate the data, it does not distinguish between such models, i.e., all linear separators are "equally favored" by the logistic loss function. This is in contrast to the

SVM method which can be used to find a particularly good (i.e., large margin) linear separator. The behavior of computational schemes for LR when the dataset is separable is not so well understood in theory, though there is recent work on first-order methods [16], [31], [18], [14], [19]. One of the main goals of this paper is to formalize the "informal" computational and statistical intuitions regarding logistic regression and to provide formal results that validate (or run counter to) such intuitive statements. In particular, a natural set of questions is: can we quantify the degree of non-separability or separability of a particular dataset, and how might such a formalism inform the computational or statistical properties of solution methods for LR? Herein, we address these questions in the context of first-order methods, which are the methods of choice in the high-dimensional regime ($n \gg 0$ and/or $p \gg 0$).

In recent years, with growing volumes of data, there has been an ever-increasing need to fit accurate logistic regression models to very large datasets with $n \gg 0$ and/or $p \gg 0$. First-order methods for tackling the problem LR are appealing in this large-scale regime for several reasons. First, the computational cost per iteration of first-order methods is relatively low compared to alternatives such as Newton's method (i.e., iteratively reweighted least squares). Second, first-order methods often tend to produce statistically interesting solutions *before* they reach convergence. In particular, several first-order methods such as gradient descent and its generalizations – steepest descent and stochastic gradient descent – are known to impart implicit regularization which induces models with good out-of-sample performance on the interior of the sequence of coefficient iterates. Moreover, at least one special case of steepest descent, namely greedy coordinate descent, also imparts desirable sparsity properties along the sequence of coefficient iterates. In this paper, we focus on the method of steepest descent in an arbitrary given norm $\|\cdot\|$ – a method that encompasses both standard gradient descent and greedy coordinate descent, among others – as well as the method of stochastic gradient descent (SGD), which is particularly appealing for problems with $n \gg 0$ since SGD only needs to sample a handful of data observations at each iteration.

Towards improving our understanding of steepest descent and SGD for logistic regression, we introduce a pair of condition numbers that measure the degree of non-separability or separability of the dataset $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$. In the case when the data is not separable, we introduce a condition number DegNSEP* that precisely quantifies the degree of non-separability of the dataset, namely datasets that are "more non-separable" have larger values of DegNSEP*. We then show that DegNSEP* naturally informs the computational guarantees of steepest descent in the sense that the guarantees improve when DegNSEP* is larger. Furthermore, we extend the definition of DegNSEP* to measure the degree of non-separability of an arbitrary distribution over the data, which allows us to analyze the role of DegNSEP* in the computational guarantees of SGD in full generality. In particular, we demonstrate that better convergence bounds – and therefore less data samples from a statistical learning point of view – are achieved when DegNSEP* is larger. In the case of separable data, we use the well-known concept of the margin [9], which we refer to as DegSEP*, to precisely quantify the degree of separability of the dataset, whereby datasets that are "more separable" have larger values of DegSEP*. We then develop computational guarantees for both ℓ_2 steepest descent and SGD that naturally depend on DegSEP* and that demonstrate convergence towards approximate-maximum-margin solutions. We also demonstrate that both DegNSEP* and DegSEP* may also be interpreted through the lens of data perturbations and the "distance to ill-posedness" introduced by Renegar [25].

There has recently been other research activity on the analysis of the performance of steepest descent and SGD for solving the logistic regression optimization problem (2). Ji and Telgarsky [16], Soudry et al. [31], Nacson et al. [18], Gunasekar et al. [14], and Nacson et al. [19] analyze the convergence properties of steepest descent and/or SGD in terms of the loss function values and iterate values when the problem instance is separable (or partially separable, this latter case not having received any previous attention that we are aware of). We discuss the results in these papers relative to ours in the relevant sections herein.

The paper is organized as follows. In Section 2 we present a pair of condition numbers for logistic regression instances, one for instances that are non-separable and another for instances that are separable. In Section 3 we examine the steepest descent algorithm (in any given norm) and show how the degree of non-separability naturally informs computational guarantees of steepest descent. In the separable case, we develop computational guarantees for ℓ_2 steepest descent that are informed by the degree of separability and show convergence towards an approximate-maximum-margin solution. Expanding on Bach [2,3], we also show how the degree of non-separability enters into the analysis of linear convergence of steepest descent (without needing strong convexity). In Section 4 we examine the stochastic gradient descent (SGD) method and we show how our condition numbers inform the computational guarantees of SGD. In the non-separable case, we show how the degree of non-separability informs standard guarantees of SGD as well as the adaptive guarantee developed by Bach in [3]. In the separable case, we develop computational guarantees for SGD that are informed by the degree of separability and show convergence in probability towards an approximate-maximum-margin solution.

1.1 Notation

For a vector $x \in \mathbb{R}^p$, x_j denotes the j^{th} coordinate; we use superscripts to index vectors in a sequence $\{x^k\}$. Let e_j denote the j^{th} unit vector in $\mathbb{R}^p$, and let $e = (1, \ldots, 1)$. We will make use of a generic given norm $\|\cdot\|$ on $\mathbb{R}^p$ as well as the ℓ_q norm denoted by $\|\cdot\|_q$ with unit ball B_q for $q \in [1, \infty]$. For the given norm $\|\cdot\|$, $\|\cdot\|_*$ denotes the dual norm defined by $\|s\|_* = \max\limits_{x:\|x\|\leq 1} s^T x$. Let $\text{Dist}(v, S) := \min_{x \in S} \|x - v\|$ denote the distance from a point v to a set S. Let $\|v\|_0$ denote the number of non-zero coefficients of the vector v.

For $A \in \mathbb{R}^{n \times p}$, let $\|A\|_{\cdot,q} := \max\limits_{x:\|x\|\leq 1} \|Ax\|_q$ be the operator norm using the given norm $\|\cdot\|$, and let $\|A\|_{q_1,q_2} := \max\limits_{x:\|x\|_{q_1}\leq 1} \|Ax\|_{q_2}$ be the operator norm where the given norm is the ℓ_{q_1} norm. Furthermore, let $\text{null}(A) := \{x \in \mathbb{R}^p : Ax = 0\}$ denote the null space of A. For a symmetric matrix M, we write "$M \succeq 0$" to denote that M is positive semidefinite, "$M \succ 0$" to denote that M is positive definite, and let $\lambda_{\min}(M)$ denote the smallest eigenvalue of M.

For a scalar α, $\text{sgn}(\alpha)$ denotes the sign of α, and α^+, α^- denote the positive and negative parts of α, respectively. The notation "$\tilde{v} \leftarrow \arg\max\limits_{v \in S}\{f(v)\}$" denotes assigning $\tilde{v}$ to be any optimal solution of the problem $\max\limits_{v \in S}\{f(v)\}$.

2 Logistic Regression, and a Pair of Condition Numbers for Non-Separable and Separable Training Data

Let us review the setting and notation of logistic regression and the basic properties of the optimization problem LR. Recall that we have n observed training data points $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_i \in \mathbb{R}^p$ is the vector of feature values and $y_i \in \{-1, 1\}$ is the (binary) class of observation i, for $i = 1, \ldots, n$. Let $\mathbf{X}$ be the matrix whose i^{th} row is $\mathbf{x}_i$, for $i = 1, \ldots, n$. The well-known logistic loss function is $L_n(\cdot) : \mathbb{R}^p \to \mathbb{R}$ defined by:

$$L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln\left(1 + \exp\left(-y_i \beta^T \mathbf{x}_i\right)\right) \ , \tag{3}$$

where $\beta \in \mathbb{R}^p$. Throughout the paper, we denote the univariate logistic loss function by $\ell(t) := \ln(1 + \exp(-t))$, the gradient of $L_n(\cdot)$ by $\nabla L_n(\cdot)$, and the Hessian of $L_n(\beta)$ by $H(\beta)$.

As mentioned previously, LR, the problem of minimizing the logistic loss function $L_n(\cdot)$ over $\beta \in \mathbb{R}^p$ arises from maximum likelihood estimation for the model (1). Note that $L_n(\beta) > 0$ for all β, hence $L_n^* \geq 0$ and is therefore finite. Unlike, for example, the least-squares loss function for linear regression, it is not clear *a priori* if the logistic regression problem LR has an optimal solution. Indeed, in the case when the data is separable, i.e., there exists a vector $\beta \in \mathbb{R}^p$ satisfying $y_i \beta^T \mathbf{x}_i > 0$ for $i = 1, \ldots, n$, then $L_n(\theta\beta) \to 0 = L_n^*$ as $\theta \to +\infty$, and hence LR does not attain its optimum.

In order to better understand the behavior of the logistic regression problem in general as well as the behavior of the steepest descent and SGD for logistic regression, we now develop a pair of condition numbers that measure the degree of non-separability and separability of the dataset $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)$. In particular, we show in this section that the behavior of LR in terms of existence of optima, as well as the existence of low-norm optima, can be characterized in terms of these condition numbers.

2.1 Non-Separable Data

Let us first consider the case of non-separable data. We say that observation i is *correctly classified* by β if $y_i \beta^T \mathbf{x}_i > 0$, and is *misclassified* by β if $y_i \beta^T \mathbf{x}_i \leq 0$. Letting Y denote the diagonal matrix whose i^{th} component is y_i, then with this notation observation i is correctly classified or misclassified by β if $(Y\mathbf{X}\beta)_i > 0$ or $(Y\mathbf{X}\beta)_i \leq 0$, respectively. We say that the training data is *non-separable* if there is no β that correctly classifies every observation, i.e., there is no β that satisfies $Y\mathbf{X}\beta > 0$, and in this case we write "$(\mathbf{X}, y)$ is not separable" to denote that the data $(\mathbf{X}, y)$ are not separable.

Clearly, some non-separable datasets might be "more non-separable" than others, so let us now introduce a way to measure the extent to which the dataset is non-separable. Let $\| \cdot \|$ denote the given norm on the space $\mathbb{R}^p$ of model coefficients β. We define the degree of non-separability of the training data (with respect to the norm $\| \cdot \|$) to be:

$$\begin{array}{rll} \text{DegNSEP}^* := & \min\limits_{\beta \in \mathbb{R}^p} & \frac{1}{n}\sum_{i=1}^n [y_i\beta^T\mathbf{x}_i]^- \\ & & \\ & \text{s.t.} & \|\beta\| = 1 \ , \end{array} \tag{4}$$

which states that DegNSEP* is the smallest (over all normalized models β) average misclassification error of the model β over the n observations. We emphasize that here and for the remainder of this paper that the norm $\|\cdot\|$ on the space of model coefficients $\mathbb{R}^p$ is given and fixed. (In Sections 3.4 and 4 we will require the norm $\|\cdot\|$ to be the ℓ_2 norm, but otherwise it is generic.)

Define $\beta^0 := 0 \in \mathbb{R}^p$. Noticing that $L_n(\beta^0) = \ln(2)$, the following object measures the maximum distance from any β in the level set of β^0 – namely $\{\beta \in \mathbb{R}^p : L_n(\beta) \le \ln(2)\}$ – to the set of optimal solutions of LR:

$$\text{Dist}_0 = \max_{\beta : L_n(\beta) \le \ln(2)} \left\{ \min_{\beta^* : L_n(\beta^*) = L_n^*} \|\beta - \beta^*\| \right\} \ . \tag{5}$$

The following proposition shows that the behavior of the logistic regression problem LR can be characterized in terms of the degree of non-separability DegNSEP*.

Proposition 2.1. *If* DegNSEP$^* > 0$*, then:*

(i) there is a unique optimal solution β^* of the logistic regression problem LR,

(ii) $H(\beta^*) \succ 0$,

(iii) $\|\beta^*\| \ \le \ \dfrac{L_n^*}{\text{DegNSEP}^*} \ \le \ \dfrac{\ln(2)}{\text{DegNSEP}^*}$ *, and*

(iv) $\text{Dist}_0 \ \le \ \dfrac{\ln(2) + L_n^*}{\text{DegNSEP}^*} \ \le \ \dfrac{2\ln(2)}{\text{DegNSEP}^*}$.

Proof: Suppose that DegNSEP$^* > 0$. Notice that this implies that null$(\mathbf{X}) = \{0\}$ since if this is not the case, then there exists $\bar{\beta} \in$ null$(\mathbf{X})$ with $\|\bar{\beta}\| = 1$ which implies that DegNSEP$^* = 0$. For any $\beta \in \mathbb{R}^p$, a simple calculation yields that $H(\beta) = \frac{1}{n}\mathbf{X}^T G \mathbf{X}$ where G is the $n \times n$ diagonal matrix whose i^{th} component is $G_{ii} = \ell''(y_i\beta^T\mathbf{x}_i) = \frac{\exp(y_i\beta^T\mathbf{x}_i)}{(\exp(y_i\beta^T\mathbf{x}_i)+1)^2} > 0$. Therefore, for any $\beta \in \mathbb{R}^p$, we have that null$(H(\beta)) =$ null$(\mathbf{X}) = \{0\}$, and it then follows that $H(\beta) \succ 0$. This implies that $L_n(\cdot)$ is globally strictly convex.

Notice that $\ln(1 + e^{-t}) \ge t^-$ for any t, and hence the objective function of (4) satisfies $L_n(\beta) \ge \frac{1}{n}\sum_{i=1}^n [y_i\beta^T\mathbf{x}_i]^-$ for any β. It then follows from (4) that $L_n(\beta) \ge \text{DegNSEP}^*\|\beta\|$ for all $\beta \in \mathbb{R}^p$, which rearranges to:

$$\|\beta\| \le \frac{L_n(\beta)}{\text{DegNSEP}^*} \qquad \text{for all } \beta \in \mathbb{R}^p \ . \tag{6}$$

Notice that $L_n(0) = \ln(2)$, and therefore the level set $\{\beta \in \mathbb{R}^p : L_n(\beta) \le \ln(2)\} \subset \{\beta \in \mathbb{R}^p : \|\beta\| \le \ln(2)/\text{DegNSEP}^*\}$ and hence is a nonempty compact set. It then follows from the continuity of $L_n(\cdot)$ in conjunction with the Weierstrass Theorem that LR attains its optimum. Since $L_n(\cdot)$ is strictly convex there is a unique optimal solution β^*, which proves *(i)*. By the previous discussion we have that $H(\beta^*) \succ 0$, which proves *(ii)*, and it follows from (6) that $\|\beta^*\| \le \frac{L_n^*}{\text{DegNSEP}^*}$, which proves *(iii)*. If β satisfies $L_n(\beta) \le \ln(2)$ it follows that

$$\|\beta - \beta^*\| \le \|\beta\| + \|\beta^*\| \le \frac{\ln(2)}{\text{DegNSEP}^*} + \frac{L_n^*}{\text{DegNSEP}^*} \ ,$$

which then implies *(iv)*. □

Part *(iii)* of Proposition 2.1 states that the norm of the unique optimal solution of the logistic regression problem LR is bounded inversely proportional to DegNSEP*, and part *(iv)* of Proposition 2.1 presents a similar bound on Dist_0. Part *(ii)* of Proposition 2.1 states that $H(\beta^*)$ is positive definite, i.e., that the logistic loss function is locally strongly convex at the optimum β^*. We can measure the degree of local strong convexity by defining, for any symmetric positive semidefinite matrix M, the local strong convexity constant of M (with respect to the norm $\|\cdot\|$) to be:

$$\begin{aligned} \nu^*(M) := \quad \min_{\beta \in \mathbb{R}^p} \quad & \beta^T M \beta \\ \text{s.t.} \quad & \|\beta\| = 1 \ . \end{aligned} \tag{7}$$

Part *(ii)* of Proposition 2.1 immediately implies that $\nu^*(H(\beta^*)) > 0$. Notice that when the norm $\|\cdot\|$ is the ℓ_2 norm, then $\nu^*(M) = \lambda_{\min}(M)$. It also follows from norm equivalence that there exist constants C_1 and C_2 with $0 < C_1 \le C_2$ such that $C_1\lambda_{\min}(M) \le \nu^*(M) \le C_2\lambda_{\min}(M)$ for all symmetric positive semidefinite matrices M. Proposition 2.2 below provides a lower bound on $\nu^*(H(\beta^*))$ that depends entirely on DegNSEP* and magnitude properties of $\mathbf{X}$.

Proposition 2.2. *If* DegNSEP$^* > 0$*, then:*

$$\nu^*(H(\beta^*)) \ \ge \ \tfrac{1}{4n}\nu^*(\mathbf{X}^T\mathbf{X}) \exp\left(-\frac{\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right) > 0 \ ,$$

where β^* is the unique optimal solution of LR.

Proof: Recall that $H(\beta^*) = \frac{1}{n}\mathbf{X}^T G \mathbf{X}$ where G is the $n \times n$ diagonal matrix whose i^{th} component $G_{ii} = \ell''(y_i(\beta^*)^T\mathbf{x}_i)$ satisfies

$$G_{ii} = \frac{\exp(y_i(\beta^*)^T\mathbf{x}_i)}{(\exp(y_i(\beta^*)^T\mathbf{x}_i)+1)^2} \ge \tfrac{1}{4}\exp(-|y_i(\beta^*)^T\mathbf{x}_i|) \ge \tfrac{1}{4}\exp(-\|\mathbf{x}_i\|_*\|\beta^*\|) \ge \tfrac{1}{4}\exp\left(-\tfrac{\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right) \ ,$$

where the final inequality uses part *(iii)* of Proposition 2.1 and $\|\mathbf{X}\|_{\cdot,\infty} = \max\limits_{i \in \{1,\ldots,n\}} \|\mathbf{x}_i\|_*$. Therefore, for any $\beta \in \mathbb{R}^p$, we have

$$\beta^T H(\beta^*)\beta = \tfrac{1}{n}(\mathbf{X}\beta)^T G(\mathbf{X}\beta) \ge \tfrac{1}{4n}\exp\left(-\tfrac{\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right)\beta^T(\mathbf{X}^T\mathbf{X})\beta \ ,$$

and taking the minimum of both sides of the above inequality over all β satisfying $\|\beta\| = 1$ yields the desired result. □

It turns out that DegNSEP* can also be interpreted as the minimal *perturbation* of the data for which the perturbed problem data is separable. This will be given a precise definition and proof in Section 2.3.

2.2 Separable Data

We say that the training data is *separable* if there exists β for which $Y\mathbf{X}\beta > 0$, i.e., there is a model β that correctly classifies every observation, and we write "$(\mathbf{X}, y)$ is separable" to denote that the data $(\mathbf{X}, y)$ are separable. Akin to the case of non-separable data, some separable datasets might be "more separable" than others. Employing the standard lens of statistical machine learning [15], we can measure the degree of separability using the well-known concept of the "margin" [9], which we now review for completeness. Suppose that $(\mathbf{X}, y)$ is separable, and let β be a model that correctly classifies all observations, namely $Y\mathbf{X}\beta > 0$. Then the *margin* of β is denoted by $\rho(\beta)$ and is defined to be the least classification value of β over all observations:

$$\rho(\beta) := \min_{i\in\{1,\ldots,n\}} [y_i\beta^T\mathbf{x}_i] \ .$$

We define the degree of separability of the data to be the maximum margin over all normalized models β, namely:

$$\begin{array}{rll} \mathrm{DegSEP}^* := & \max\limits_{\beta} & \rho(\beta) \\ & \text{s.t.} & \|\beta\| \le 1 \ . \end{array} \tag{8}$$

Proposition 2.3. *If* $(\mathbf{X}, y)$ *is separable, then* $\mathrm{DegSEP}^* > 0$*,* $L_n^* = 0$*, and LR does not attain its optimum.*

Proof: If $(\mathbf{X}, y)$ is separable, it follows from the definition of the margin that $\mathrm{DegSEP}^* > 0$, and there exists a vector $\bar\beta \in \mathbb{R}^p$ satisfying $\|\bar\beta\| \le 1$ and $y_i\bar\beta^T\mathbf{x}_i \ge \mathrm{DegSEP}^* > 0$ for $i = 1, \ldots, n$, whereby $L_n(\theta\bar\beta) \to 0$ as $\theta \to +\infty$. Since $L_n(\beta) > 0$ for any β, it also follows that $L_n^* \ge 0$, and hence $L_n^* = 0$ and LR does not attain its optimum. □

The following lemma relates the margin function $\rho(\cdot)$ to the gradient of the logistic loss function. In Lemma 2.1 and elsewhere in this paper we use the convention $\ln(a) = -\infty$ for $a \le 0$, i.e., $\ln(\cdot)$ is an extended-real-valued concave function.

Lemma 2.1. *Suppose that the data* $(\mathbf{X}, y)$ *is separable, i.e.,* $\mathrm{DegSEP}^* > 0$*. Then for any* $\beta \in \mathbb{R}^p$ *it holds that:*

$$\rho(\beta) \ \ge \ \ln\left(\tfrac{\mathrm{DegSEP}^*}{n\|\nabla L_n(\beta)\|_*} - 1\right) \ .$$

Proof: Recall that

$$\rho(\beta) := \min_{i\in\{1,\ldots,n\}} [y_i\beta^T\mathbf{x}_i] = \min_{i\in\{1,\ldots,n\}} (Y\mathbf{X}\beta)_i = \min_{w\in\Delta_n} w^TY\mathbf{X}\beta \ ,$$

where $\Delta_n := \{w \in \mathbb{R}^n : e^Tw = 1, \ w \ge 0\}$. By minimax strong duality it holds that:

$$\mathrm{DegSEP}^* := \max_{\beta:\|\beta\|\le 1} \ \rho(\beta) \ = \ \min_{w\in\Delta_n} \|\mathbf{X}^TYw\|_* \ . \tag{9}$$

Straightforward calculation yields that the gradient of the logistic loss function satisfies $\nabla L_n(\beta) = \frac{1}{n}\mathbf{X}^T Y w^*(\beta)$ where $w^*(\beta) \in \mathbb{R}^n$ is the vector defined component-wise by:

$$w^*(\beta)_i = \frac{1}{1+\exp(y_i \beta^T \mathbf{x}_i)} \quad , \quad i = 1, \ldots, n \ . \tag{10}$$

Since $w^*(\beta) > 0$ it follows that $w^*(\beta)/\|w^*(\beta)\|_1$ is a feasible solution of the right-most optimization problem in (9), which implies that $\frac{\|\mathbf{X}^T Y w^*(\beta)\|_*}{\|w^*(\beta)\|_1} \geq \mathrm{DegSEP}^*$. Thus, since $\mathrm{DegSEP}^* > 0$, it holds that

$$\|w^*(\beta)\|_1 \leq \frac{\|\mathbf{X}^T Y w^*(\beta)\|_*}{\mathrm{DegSEP}^*} = \frac{n\|\nabla L_n(\beta)\|_*}{\mathrm{DegSEP}^*} \ .$$

In particular, for each $i = 1, \ldots, n$, it holds that $w^*(\beta)_i \leq \frac{n\|\nabla L_n(\beta)\|_*}{\mathrm{DegSEP}^*}$, which, after simple arithmetic manipulation using (10), is equivalent to:

$$y_i \beta^T \mathbf{x}_i \ \geq \ \ln\left(\frac{\mathrm{DegSEP}^*}{n\|\nabla L_n(\beta)\|_*} - 1\right) \ , \ i = 1, \ldots, n \ . \tag{11}$$

Since (11) holds for all $i = 1, \ldots, n$, the result is proved. □

It turns out that DegSEP* can also be interpreted as the minimal perturbation of the data for which the perturbed problem data is non-separable. This is developed in the following subsection.

2.3 Data-Perturbation Interpretations of DegNSEP* and DegSEP*

In this subsection we show that both DegNSEP* and DegSEP* can be interpreted through the lens of data perturbations that alter the status of the dataset – from non-separable to separable, or *vice versa*. Let us view the feature data matrix $\mathbf{X}$ as a linear operator, and recall the operator norm notation $\|\mathbf{X}\|_{\cdot,q} := \max\limits_{\beta:\|\beta\|\leq 1} \|\mathbf{X}\beta\|_q$ on the space $\mathbb{R}^{n\times p}$.

Define:

$$\begin{array}{rll} \mathrm{PertSEP}^* := & \inf\limits_{\Delta\mathbf{X}} & \frac{1}{n}\|\Delta\mathbf{X}\|_{\cdot,1} \\ & \text{s.t.} & (\mathbf{X}+\Delta\mathbf{X}, y) \text{ is separable} \ . \end{array} \tag{12}$$

Then PertSEP* is the smallest (or more precisely, the infimum thereof) perturbation $\Delta\mathbf{X}$ of the feature data $\mathbf{X}$ which will render the perturbed problem instance $(\mathbf{X}+\Delta\mathbf{X}, y)$ separable. Here the size of the perturbation is measured using the (scaled) operator norm $\frac{1}{n}\|\cdot\|_{\cdot,1}$. Clearly $\mathrm{PertSEP}^* = 0$ if $(\mathbf{X}, y)$ is separable. If $\mathrm{PertSEP}^* > 0$, then $(\mathbf{X}, y)$ is not separable; and the smaller PertSEP* is, the closer the data is to being separable. Notice that $\|\Delta\mathbf{X}\|_{\cdot,1}$ scales proportional to n, which is counteracted by dividing by n in the objective function of (12). The following result shows that the condition number DegNSEP* introduced and used in Section 2.1 can be alternatively interpreted as the smallest data perturbation for which the perturbed data $(\mathbf{X}+\Delta\mathbf{X}, y)$ is separable.

Proposition 2.4. *For any dataset* $(\mathbf{X}, y)$ *it holds that* $\mathit{DegNSEP}^* = \mathit{PertSEP}^*$.

Proof: See Appendix A.1. □

Let us also define:

$$\begin{aligned} \text{PertNSEP}^* := \ & \inf_{\Delta \mathbf{X}} \quad \|\Delta \mathbf{X}\|_{\cdot,\infty} \\ & \text{s.t.} \quad (\mathbf{X} + \Delta \mathbf{X}, y) \text{ is non-separable .} \end{aligned} \tag{13}$$

Then PertNSEP* is the smallest perturbation $\Delta\mathbf{X}$ of the feature data $\mathbf{X}$ which will render the perturbed problem instance $(\mathbf{X} + \Delta\mathbf{X}, y)$ non-separable. Here the size of the perturbation is measured using the operator norm $\| \cdot \|_{\cdot,\infty}$. Clearly PertNSEP$^* = 0$ if $(\mathbf{X}, y)$ is not separable. If PertNSEP$^* > 0$, then $(\mathbf{X}, y)$ is separable; and the smaller PertNSEP* is, the closer the data is to being non-separable. Notice that we use a different operator norm in the definition of PertNSEP* than that used in the definition of PertSEP*. The following result shows that the condition number DegSEP* introduced and used in Section 2.2 can be alternatively interpreted as the smallest data perturbation for which the perturbed data $(\mathbf{X} + \Delta\mathbf{X}, y)$ is non-separable.

Proposition 2.5. *For any dataset* $(\mathbf{X}, y)$ *it holds that DegSEP* = PertNSEP*.*

Proof: See Appendix A.1 as well. □

Any given dataset $(\mathbf{X}, y)$ is either non-separable (in which case DegSEP$^* = 0$) or is separable (in which case DegNSEP$^* = 0$). Borrowing from the lexicon of Renegar [25], we may call the dataset $(\mathbf{X}, y)$ "ill-posed" if both DegNSEP$^* = 0$ and DegSEP$^* = 0$, in which case the dataset is non-separable but there is an arbitrarily small perturbation of the data that renders the perturbed dataset separable. It can easily be verified that an example of such a dataset is:

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & -2 & 3 \\ 2 & 1 & -3 \end{pmatrix} \ , \quad y = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \ .$$

3 Informing Standard Deterministic First-Order Solution Methods for Logistic Regression

In this section we show how the two condition numbers DegNSEP* and DegSEP* inform the computational properties and guarantees of a standard deterministic first-order solution method for logistic regression, namely the steepest descent method in any given norm $\| \cdot \|$. After briefly reviewing steepest descent in the setting of smooth convex optimization in Section 3.1, we examine steepest descent as applied to logistic regression in two cases: in Section 3.2 we consider the case of non-separable data and examine steepest descent for any given norm, and in Section 3.4 we consider the case of separable data and examine steepest descent for the ℓ_2 norm.

3.1 Brief Review of Steepest Descent

We review the steepest descent method for solving the unconstrained optimization problem:

$$\text{(P):} \quad f^* := \min_{x \in \mathbb{R}^p} \ f(x) \ , \tag{14}$$

where $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is a differentiable convex function, and we assume that f^* is finite but it is not necessarily attained. Let $\| \cdot \|$ be the norm on the variables $x \in \mathbb{R}^p$. At a given iterate $\bar{x}$, the steepest descent direction is defined to be the negative of the normalized direction $\bar{d}$ that maximizes $\nabla f(\bar{x})^T d$, namely: $\bar{d} \leftarrow \arg\max_d \{\nabla f(\bar{x})^T d : \|d\| \le 1\}$. The formal statement of steepest descent in the norm $\| \cdot \|$ is presented in Algorithm 1.

Algorithm 1 Steepest Descent in the norm $\| \cdot \|$

```
Initialize at $x^0 \in \mathbb{R}^p$, $k \leftarrow 0$

At iteration $k$:
1. Compute $\nabla f(x^k)$
2. Compute $d^k \leftarrow \arg\max_d \{\nabla f(x^k)^T d : \|d\| \le 1\}$
3. Choose $\alpha_k \ge 0$ and set:
   $x^{k+1} \leftarrow x^k - \alpha_k \cdot d^k$
```

To the best of our knowledge, there is no general computational guarantee associated with steepest descent for an arbitrary norm without additional assumptions such as bounded optimal solutions, strong convexity of the function, and/or maximum distances of the starting points and/or the iterates from the set of optimal solutions (or the set of near-optimal solutions). In the Euclidean norm case ($\| \cdot \| = \| \cdot \|_2$), steepest descent specifies to the classical gradient descent algorithm, see [24] and [22]. In the particular case when $\| \cdot \|$ is the ℓ_1 norm $\| \cdot \|_1$, steepest descent specifies to the greedy coordinate descent method. For works related to greedy coordinate descent, including block-coordinate descent, cyclic and randomized coordinate descent and variations thereof, see for example Nesterov [21], Richtárik and Takáč [26], Schmidt and Friedlander [28], and Beck and Tetruashvili [5].

Recall that $f(\cdot)$ is L-smooth with respect to the norm $\| \cdot \|$ if $f(\cdot)$ is differentiable and satisfies:

$$\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\| \quad \text{for all } x, y \in \mathbb{R}^p \ , \tag{15}$$

where $\| \cdot \|_*$ is the dual norm of $\| \cdot \|$.

Let $\mathcal{S}_0$ denote the level set of the initial iterate x^0, namely $\mathcal{S}_0 := \{x \in \mathbb{R}^p : f(x) \le f(x^0)\}$, and let $\mathcal{S}^*$ denote the set of optimal solutions of (14), i.e., $\mathcal{S}^* := \{x \in \mathbb{R}^p : f(x) = f^*\}$. Then let Dist_0 denote the largest distance of points in $\mathcal{S}_0$ to the set of optimal solutions $\mathcal{S}^*$:

$$\text{Dist}_0 := \max_{x \in \mathcal{S}_0} \min_{x^* \in \mathcal{S}^*} \|x - x^*\| \ . \tag{16}$$

The following computational guarantees for steepest descent are an amended and extended version of mostly well-known results about traditional gradient descent and greedy coordinate descent, see for example Nesterov [22] and Beck and Tetruashvili [5].

Theorem 3.1. (Computational Guarantees for Steepest Descent in the norm $\|\cdot\|$) *Let $\{x^k\}$ be generated according to the steepest descent method (Algorithm 1) using the step-size sequence $\{\alpha_k\}$ chosen using the "greedy" rule:*

$$\alpha_k = \frac{\|\nabla f(x^k)\|_*}{L} \quad \text{for all } k \geq 0 \ . \tag{17}$$

If $x^0 \notin \mathcal{S}^*$, then it holds for all $k \geq 0$ that:

(i) (optimality gap): $f(x^k) - f^* \ \leq \ \dfrac{2L(\mathrm{Dist}_0)^2}{\hat{K}^0 + k} \ < \ \dfrac{2L(\mathrm{Dist}_0)^2}{k}$

and

(ii) (gradient bound I): $\|\nabla f(x^k)\|_* \ \leq \ \sqrt{2L(f(x^k) - f^*)} \ \leq \ \dfrac{2L\mathrm{Dist}_0}{\sqrt{\hat{K}^0 + k}}$

where $\hat{K}^0 := \dfrac{2L(\mathrm{Dist}_0)^2}{f(x^0) - f^*}$. Furthermore,

(iii) (norm bound): $\|x^k - x^0\| \ \leq \ \sqrt{k}\sqrt{\dfrac{2(f(x^0) - f^*)}{L}}$, and

(iv) (gradient bound II): there exists $i \leq k$ for which $\|\nabla f(x^i)\|_* \ \leq \ \sqrt{\dfrac{2L(f(x^0) - f^*)}{k+1}}$,

where the two inequalities in (i) and the second inequality in (ii) are only relevant when Dist$_0$ *is finite.* □

For completeness, a self-contained proof of Theorem 3.1 is given in Appendix A.2.

Remark 3.1. If an exact line-search is used instead of the step-size rule (17), then all of the results in Theorem 3.1 remain valid except for the norm bound in item *(iii)*. This follows easily from the structure of the proof in Appendix A.2.

Remark 3.2. In the case of the ℓ_2 norm, Dist$_0$ can be replaced by the typically much small quantity Dist$(x^0, \mathcal{S}^*)$ in items *(i)* and *(ii)* of the Theorem 3.1.

3.2 Informing Steepest Descent for Solving Logistic Regression in the Non-Separable Case

Here we show how the condition numbers DegNSEP* and DegSEP* inform computational guarantees for steepest descent applied to the logistic regression optimization problem (2). In order to apply Theorem 3.1 to the setting of logistic regression, we use the following smoothness property of the logistic loss function.

Proposition 3.1. (Lipschitz smoothness of the logistic loss function) *The logistic loss function $L_n(\cdot)$ is $L = \frac{1}{4n}\|\mathbf{X}\|_{\cdot,2}^2$-smooth with respect to the given norm $\|\cdot\|$ on $\mathbb{R}^p$.*

Proof: See Appendix A.3. □

Theorem 3.2. (Steepest Descent for Logistic Regression in the norm $\|\cdot\|$: Non-Separable Case) *Suppose that steepest descent (Algorithm 1) is initialized at $\beta^0 := 0$, and is implemented using the step-size rule:*

$$\alpha_k := \frac{4n\|\nabla L_n(\beta^k)\|_*}{\|\mathbf{X}\|_{\cdot,2}^2} \quad \textit{for all } k \geq 0 \ . \tag{18}$$

When the data is non-separable, for all $k \geq 0$ it holds that:

(i) (training error):

$$L_n(\beta^k) - L_n^* \ \leq \ \frac{1}{\dfrac{1}{\ln(2) - L_n^*} + \dfrac{k \cdot n \cdot (\mathrm{DegNSEP}^*)^2}{2\|\mathbf{X}\|_{\cdot,2}^2(\ln(2))^2}} \ < \ \frac{2\|\mathbf{X}\|_{\cdot,2}^2(\ln(2))^2}{k \cdot n \cdot (\mathrm{DegNSEP}^*)^2} \ ,$$

(ii) (shrinkage): $\|\beta^k\| \ \leq \ \sqrt{k}\left(\frac{1}{\|\mathbf{X}\|_{\cdot,2}}\right)\sqrt{8n(\ln(2) - L_n^*)} \ \leq \ \sqrt{k}\left(\frac{1}{\|\mathbf{X}\|_{\cdot,2}}\right)\sqrt{8n\ln(2)}$ *, and*

(iii) (gradient bound): $\|\nabla L_n(\beta^k)\|_* \ \leq \ \|\mathbf{X}\|_{\cdot,2}\sqrt{\frac{(L_n(\beta^k) - L_n^*)}{2n}} \ \leq \ \dfrac{\|\mathbf{X}\|_{\cdot,2}^2\ln(2)}{\sqrt{k}\cdot n \cdot \mathrm{DegNSEP}^*}$.

Proof: These results are a straightforward application of Theorem 3.1, Proposition 3.1, and Proposition 2.1. Parts *(i)* and *(ii)* of Theorem 3.1 present computational guarantees for the steepest descent method for the step-size rule (17) in terms of the initial objective function value (which in this case is $L_n(\beta^0) = L_n(0) = \ln(2)$), the Lipschitz constant L, and the distance measure Dist_0 defined in (16). From Proposition 3.1 we can take the Lipschitz constant L of the gradient of the logistic loss function $L_n(\cdot)$ to be $L = \frac{1}{4n}\|\mathbf{X}\|_{\cdot,2}^2$. And from Proposition 2.1 we have that $\mathrm{Dist}_0 \leq \frac{2\ln(2)}{\mathrm{DegNSEP}^*}$ in the case when the data is non-separable. Substituting these values into the step-size formula (17) and utilizing parts *(i)* and *(ii)* of Theorem 3.1 yields precisely the step-size rule (18) and the computational guarantees in parts *(i)* and *(iii)* of the present theorem. Also, substituting the bound on L into part *(iii)* of Theorem 3.1 yields part *(ii)* of the present theorem. □

Notice the manner in which $\mathrm{DegNSEP}^*$ informs the computational guarantees in Theorem 3.2. The training error bound scales like $O(1/(\mathrm{DegNSEP}^*)^2)$, and so the computational guarantee on the training error is smaller for datasets with larger values of $\mathrm{DegNSEP}^*$. Also note that the training error bound is invariant under constant re-scaling of the data – since rescaling all observations by a constant γ will rescale both $\mathrm{DegNSEP}^*$ and $\|\mathbf{X}\|_{\cdot,2}$ by γ and so their quotient is unaffected. Similarly, the computational guarantee on the norm of the gradient scales like $O(1/\mathrm{DegNSEP}^*)$, and is smaller for datasets with larger values of $\mathrm{DegNSEP}^*$ as well.

3.3 Linear Convergence Results

In a series of papers, Bach [2, 3], as well as Bach and Moulines [4], identified a generalized self-concordance property of the logistic loss function that has proven to be useful in analyzing the statistical and computational properties of empirical risk minimization as well as stochastic gradient descent in this setting. In particular, in the case of non-separable data, Bach demonstrates in [3] that (averaged) stochastic gradient descent is adaptive to the unknown local strong convexity of the logistic loss function at the optimum. That is, the convergence rate can be improved from

$O(1/\sqrt{k})$ to $O(1/\mu k)$ where μ is the smallest eigenvalue of the Hessian at the optimum. Since it is known that steepest descent achieves linear convergence in the case of a strongly convex objective function, it is natural to ask whether steepest descent is also adaptive to the unknown local strong convexity of the logistic loss function? Here we answer this question in the affirmative.

We show in the case of non-separable data that steepest descent achieves linear convergence with a rate of linear convergence that naturally depends on the condition number $\mathrm{DegNSEP}^*$ as well as a measure of local strong convexity at the optimum. Our results provide linear convergence guarantees both in terms of the training error gap as well as the distance to the optimal solution β^* measured in the given norm. Moreover, we show that after a certain number of iterations that scales like $O(1/(\mathrm{DegNSEP}^*)^2)$, the rate of linear convergence improves to a faster rate that is independent of $\mathrm{DegNSEP}^*$. Thus, steepest descent is generally adaptive to the local strong convexity of the logistic loss function and also achieves faster convergence in a neighborhood of an optimal solution. The proofs of our results utilize the generalized self-concordance theory of Bach [2,3], with analysis that is perhaps simpler than the case of stochastic gradient descent examined in [3].

Theorem 3.3 below presents the linear convergence results for steepest descent in the non-separable case. Recall that $\nu^*(M)$ denotes the local strong convexity constant of a symmetric positive semidefinite matrix M with respect to the given norm $\|\cdot\|$, and that part *(ii)* of Proposition 2.1 implies that $\nu^*(H(\beta^*)) > 0$ whenever $\mathrm{DegNSEP}^* > 0$.

Theorem 3.3. (Linear Convergence of Steepest Descent in the Non-Separable Case) *Suppose that Steepest Descent (Algorithm 1) is initalized at $\beta^0 := 0$, and is implemented using the step-size sequence:*

$$\alpha_k := \frac{4n\|\nabla L_n(\beta^k)\|_*}{\|\mathbf{X}\|_{\cdot,2}^2} \quad \textit{for all } k \geq 0 \ .$$

Suppose that the data is non-separable and $\mathrm{DegNSEP}^* > 0$. *Define the "slow" rate of linear convergence constant:*

$$\tau_{\text{slow}} := \left(1 - \frac{2(\mathrm{DegNSEP}^*)\nu^*(H(\beta^*))n}{(\mathrm{DegNSEP}^* + 2\ln(2)\|\mathbf{X}\|_{\cdot,\infty})\|\mathbf{X}\|_{\cdot,2}^2}\right) < 1 \ .$$

Then for all $k \geq 0$, it holds that:

(i) *(training error):* $L_n(\beta^k) - L_n^* \ \leq \ (\ln(2) - L_n^*) \cdot (\tau_{\text{slow}})^k$ *, and*

(ii) *(coefficient convergence):* $\|\beta^k - \beta^*\| \ \leq \ \left(1 + \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\mathrm{DegNSEP}^*}\right)\left(\frac{\|\mathbf{X}\|_{\cdot,2}}{\nu^*(H(\beta^*))}\right)\sqrt{\frac{\ln(2)-L_n^*}{2n}} \cdot (\tau_{\text{slow}})^{k/2}$
,

where β^* is the unique optimal solution of LR. Furthermore, define:

$$\check{K} := \left\lceil \frac{16\ln(2)^2\|\mathbf{X}\|_{\cdot,2}^4\|\mathbf{X}\|_{\cdot,\infty}^2}{9n^2(\mathrm{DegNSEP}^*)^2\nu^*(H(\beta^*))^2} \right\rceil \ ,$$

and the "fast" rate of linear convergence constant:

$$\tau_{\text{fast}} := \left(1 - \frac{\nu^*(H(\beta^*))n}{\|\mathbf{X}\|_{\cdot,2}^2}\right) < \tau_{\text{slow}} < 1 \ .$$

Then for all $k \geq \check{K}$, it holds that:

(iii) (training error): $L_n(\beta^k) - L_n^* \ \leq \ (L_n(\beta^{\check{K}}) - L_n^*) \cdot (\tau_{\text{fast}})^{k-\check{K}}$ *, and*

(iv) (coefficient convergence): $\|\beta^k - \beta^*\| \ \leq \ \frac{\|\mathbf{X}\|_{\cdot,2}}{\nu^*(H(\beta^*))}\sqrt{\frac{2(L_n(\beta^{\check{K}})-L_n^*)}{n}} \cdot (\tau_{\text{fast}})^{(k-\check{K})/2}$.

□

The proof of Theorem 3.3 is presented in Appendix A.4. As compared to results of a similar flavor for other algorithms, here we have precise guarantees for both the "slow" and "fast" rates of linear convergence as well as for the point at which the fast rate is guaranteed to "kick in." Moreover, there is a natural dependence on DegNSEP* in both the slow rate τ_{slow} as well as the iterate $\check{K}$ by which point the linear convergence rate is improved. Note that the bound on the training error provided by part *(i)* of Theorem 3.2, while sublinear, will be superior to the linear convergence bound provided by part *(i)* of Theorem 3.3 during the early iterations of steepest descent. On the other hand, the fast linear convergence bound provided by part *(iii)* of Theorem 3.3 will eventually be the superior of all three bounds when k is large enough. Recall that Proposition 2.2 states that $\nu^*(H(\beta^*))$ is bounded from below as follows:

$$\nu^*(H(\beta^*)) \ \geq \ \tfrac{1}{4n}\nu^*(\mathbf{X}^T\mathbf{X}) \exp\left(-\frac{\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right) > 0 \ ,$$

which only depends on DegNSEP* and the magnitude properties of $\mathbf{X}$. This lower bound can be leveraged to develop versions of the slow and fast linear convergence rates that depend only on DegNSEP* and the magnitude properties of $\mathbf{X}$. However (and as also noted by Bach in [3]), the exponential term in the above lower bound is quite pessimistic and in practice $\nu^*(H(\beta^*))$ tends to be not much smaller than $\frac{1}{4n}\nu^*(\mathbf{X}^T\mathbf{X})$, i.e, it is as if the exponential term above is not present.

3.4 Informing ℓ_2 Steepest Descent for Solving Logistic Regression in the Separable Case

Let us also examine the case when the data is separable. As mentioned earlier, the logistic regression optimization problem (2) is not naturally designed for the case when the data is separable due to the fact that in this case there is no optimal solution. Indeed, one would suspect that in this case most algorithms – in particular elementary algorithms such as steepest descent – would not exhibit computational guarantees of interest. However, Soudry et al. [31] and Ji and Telgarsky [16] have recently shown for the case of separable data that steepest descent for the ℓ_2 norm delivers solutions whose normalized values are approximate-maximum-margin solutions. The following theorem presents our results for ℓ_2 steepest descent in the separable case, which adds different and explicit computational guarantees in the separable case.

Theorem 3.4. **(ℓ_2 Steepest Descent for Logistic Regression: Separable Case)** *Suppose that ℓ_2 steepest descent (Algorithm 1 with the ℓ_2 norm) is initialized at $\beta^0 := 0$, and is implemented using the step-size rule:*

$$\alpha_k := \frac{2\|\nabla L_n(\beta^k)\|_2}{\|\mathbf{X}\|_{2,\infty}^2} \quad \textit{for all } k \geq 0 \ . \tag{19}$$

When the data is separable, it holds for all $k \geq 1$ that:

(i) (margin bound): there exists $i \in \{0, 1, \ldots, k\}$ for which the normalized iterate $\bar{\beta}^i := \beta^i / \|\beta^i\|_2$ satisfies

$$\rho(\bar{\beta}^i) \;\geq\; \frac{\mathrm{DegSEP}^* \cdot \ln\left(\frac{\mathrm{DegSEP}^*}{n\|\mathbf{X}\|_{2,\infty}}\sqrt{\frac{3(k+1)}{2\ln(2)}} - 1\right)}{2(\ln(k)+1)} \;, \tag{20}$$

(ii) (shrinkage): $\|\beta^k\|_2 \;\leq\; \frac{2\ln(k)}{\mathrm{DegSEP}^*} + \frac{2}{\|\mathbf{X}\|_{2,\infty}}$ *, and*

(iii) (gradient bound): $\min\limits_{i\in\{0,\ldots,k\}} \|\nabla L_n(\beta^i)\|_2 \;\leq\; \|\mathbf{X}\|_{2,\infty}\sqrt{\frac{2\ln(2)}{3(k+1)}}$. □

Note that the constant step-size value (19) in Theorem 3.4 is different from the corresponding value (18) in Theorem 3.2 (when the norm is the ℓ_2 norm). Indeed, the step-size value (19) is smaller than the step-size value (18) since $\frac{1}{n}\|\mathbf{X}\|_{2,2}^2 \leq \|\mathbf{X}\|_{2,\infty}^2$. Therefore, $\frac{1}{2}\|\mathbf{X}\|_{2,\infty}^2$ is a valid upper bound on the Lipschitz constant of the gradient and it is straightforward to develop variants of Theorems 3.2 and 3.3 for the step-size (19), wherein $\frac{1}{4n}\|\mathbf{X}\|_{2,2}^2$ would be replaced by $\frac{1}{2}\|\mathbf{X}\|_{2,\infty}^2$ in all of the bounds.

The bound in item *(i)* of Theorem 3.4 can be understood as $O(1/\ln(k))$ relative convergence to at least $\frac{\mathrm{DegSEP}^*}{4}$. To demonstrate this, consider setting $k := \lfloor \frac{2\ln(2)\Omega^2 n^2 \|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2} \rfloor$ for some parameter $\Omega \geq 2$. Then the bound in (20) becomes:

$$\rho(\bar{\beta}^i) \;\geq\; \frac{\mathrm{DegSEP}^* \cdot \ln(\Omega - 1)}{4\ln(\Omega) + 2\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + 2} \;,$$

and rearranging the above yields:

$$\begin{aligned}
\frac{\rho(\bar{\beta}^i)}{\frac{\mathrm{DegSEP}^*}{4}} \;&\geq\; \frac{\ln(\Omega - 1)}{\ln(\Omega) + \frac{1}{2}\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + \frac{1}{2}} \\
&=\; 1 \;-\; \frac{\ln\left(\frac{\Omega}{\Omega-1}\right) + \frac{1}{2}\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + \frac{1}{2}}{\ln(\Omega) + \frac{1}{2}\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + \frac{1}{2}} \\
&\geq\; 1 \;-\; \frac{\ln(2) + \frac{1}{2}\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + \frac{1}{2}}{\frac{1}{2}\ln(\Omega^2) + \frac{1}{2}\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + \frac{1}{2}} \\
&\geq\; 1 \;-\; \frac{\ln(2) + \frac{1}{2}\ln\left(\frac{2\ln(2)n^2\|\mathbf{X}\|_{2,\infty}^2}{3(\mathrm{DegSEP}^*)^2}\right) + \frac{1}{2}}{\frac{1}{2}\ln(k+1) + \frac{1}{2}} \;,
\end{aligned}$$

(where the second inequality uses $\Omega \geq 2$), and letting k grow demonstrates $O(1/\ln(k))$ convergence of the iterate margins to $\frac{\mathrm{DegSEP}^*}{4}$ or larger. Interestingly, except for the factor of 4, this result is similar to Soudry et al. [31] who show $O(1/\ln(k))$ convergence to DegSEP*, and to Ji and Telgarsky [16], whose work shows $O(\sqrt{\ln\ln(k)/\ln(k)})$ convergence to DegSEP*; however our arguments appear to be entirely different and they result in an explicit margin bound, whereas the results in [16] and [31] are less transparent. On the other hand, [16] and [31] of course prove convergence towards DegSEP^*, not $\frac{\text{DegSEP}^*}{4}$.

In order to prove Theorem 3.4, we will use the following lemma which bounds the norms of iterates of ℓ_2 steepest descent applied to the logistic regression problem (2), and which is a modified version of a result in Ji and Telgarsky [16].

Lemma 3.1. (essentially from Ji and Telgarsky [16]) *Suppose that ℓ_2 steepest descent (Algorithm 1 with the ℓ_2 norm) is initialized at $\beta^0 := 0$ using the step-size sequence $\{\alpha_k\}$. If* $\text{DegSEP}^* > 0$ *and $\alpha_k \le \frac{2\|\nabla L_n(\beta^k)\|_2}{\|\mathbf{X}\|_{2,\infty}^2}$ for all $k \ge 0$, then it holds for all $k \ge 1$ that:*

$$\|\beta^k\|_2 \ \le \ \frac{2\ln(k)}{\text{DegSEP}^*} + \frac{2}{\|\mathbf{X}\|_{2,\infty}} \ .$$

□

The proof of this lemma is presented in Appendix A.5.

Proof of Theorem 3.4: We first prove item *(iii)*. Let $i \in \{0,\ldots,k\}$. By the smoothness of the logistic loss function we have:

$$\begin{aligned} L_n(\beta^{i+1}) \ &\le \ L_n(\beta^i) + \nabla L_n(\beta^i)^T(\beta^{i+1} - \beta^i) + \tfrac{\|\mathbf{X}\|_{2,2}^2}{8n}\|\beta^{i+1} - \beta^i\|_2^2 \\ &= \ L_n(\beta^i) - \alpha_i\|\nabla L_n(\beta^i)\|_2^2 + \tfrac{\alpha_i^2\|\mathbf{X}\|_{2,2}^2}{8n}\|\nabla L_n(\beta^i)\|_2^2 \\ &= \ L_n(\beta^i) - \tfrac{2}{\|\mathbf{X}\|_{2,\infty}^2}\|\nabla L_n(\beta^i)\|_2^2 + \tfrac{\|\mathbf{X}\|_{2,2}^2}{2n\|\mathbf{X}\|_{2,\infty}^4}\|\nabla L_n(\beta^i)\|_2^2 \\ &\le \ L_n(\beta^i) - \tfrac{2}{\|\mathbf{X}\|_{2,\infty}^2}\|\nabla L_n(\beta^i)\|_2^2 + \tfrac{1}{2\|\mathbf{X}\|_{2,\infty}^2}\|\nabla L_n(\beta^i)\|_2^2 \\ &= \ L_n(\beta^i) - \tfrac{3}{2\|\mathbf{X}\|_{2,\infty}^2}\|\nabla L_n(\beta^i)\|_2^2 \ . \end{aligned}$$

Summing over all $i \in \{0,\ldots,k\}$ yields:

$$\frac{3}{2\|\mathbf{X}\|_{2,\infty}^2}\sum_{i=0}^{k}\|\nabla L_n(\beta^i)\|_2^2 \le L_n(\beta^0) - L_n(\beta^{k+1}) \le \ln(2) \ ,$$

which implies item *(iii)* after rearranging and replacing the terms in the summation by their minimum. Item *(ii)* is a restatement of Lemma 3.1. To prove item *(i)*, let i be a minimizing index in item *(iii)*, and note by item *(iii)* and Lemma 2.1 that:

$$\rho(\beta^i) \ \ge \ \ln\left(\tfrac{\text{DegSEP}^*}{n\|\nabla L_n(\beta^i)\|_2} - 1\right) \ \ge \ \ln\left(\tfrac{\text{DegSEP}^*}{n\|\mathbf{X}\|_{2,\infty}}\sqrt{\tfrac{3(k+1)}{2\ln(2)}} - 1\right) \ .$$

Combining the above with item *(ii)* and using $\text{DegSEP}^* \le \|\mathbf{X}\|_{2,\infty}$, we obtain:

$$\rho(\bar{\beta}^i) = \frac{\rho(\beta^i)}{\|\beta^i\|_2} \ \ge \ \frac{\ln\left(\tfrac{\text{DegSEP}^*}{n\|\mathbf{X}\|_{2,\infty}}\sqrt{\tfrac{3(k+1)}{2\ln(2)}} - 1\right)}{\tfrac{2\ln(i)}{\text{DegSEP}^*} + \tfrac{2}{\|\mathbf{X}\|_{2,\infty}}} \ \ge \ \frac{\ln\left(\tfrac{\text{DegSEP}^*}{n\|\mathbf{X}\|_{2,\infty}}\sqrt{\tfrac{3(k+1)}{2\ln(2)}} - 1\right)}{\tfrac{2\ln(k)}{\text{DegSEP}^*} + \tfrac{2}{\text{DegSEP}^*}} \ ,$$

and the proof then follows from rearranging terms in the above inequality. □

While Theorem 3.4 holds only for the case of ℓ_2 steepest descent, it is also possible to derive a much weaker result for any given norm $\|\cdot\|$ by employing the $O(\sqrt{k})$ iterate norm bound in item *(ii)* of Theorem 3.2 instead of the $O(\ln(k)/\mathrm{DegSEP}^*)$ bound given by Lemma 3.1. We also mention that Gunasekar et al. [14] show that $\lim_{k\to\infty}\rho(\bar{\beta}^k) = \mathrm{DegSEP}^*$ for steepest descent in an arbitrary norm, although there is no analysis of the rate of convergence. Of course, it would be very interesting to see if the tools used to prove these various results can be combined to yield stronger convergence guarantees about the margin.

4 Informing Stochastic Gradient Descent for Logistic Regression

In this section, we examine the role of the condition numbers DegNSEP* and DegSEP* in the computational and statistical properties of the stochastic gradient descent (SGD) method applied to logistic regression. Throughout this section, we consider a more general version of the logistic regression problem LR that replaces the empirical average in (2) with an arbitrary distribution. Let $\mathcal{D}$ denote an arbitrary distribution on the data $(\mathbf{x}, y)$. We consider the following version of the logistic regression problem:

$$\begin{array}{rlrl} \mathrm{LR}_{\mathcal{D}} \ : \quad L^*_{\mathcal{D}} := & \min\limits_{\beta} & L_{\mathcal{D}}(\beta) \ := \ \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ln\left(1+\exp\left(-y\beta^T\mathbf{x}\right)\right)\right] \\ & \text{s.t.} & \beta \in \mathbb{R}^p \ . \end{array} \tag{21}$$

Note that there are two important special cases of $\mathrm{LR}_{\mathcal{D}}$. When $\mathcal{D}$ is the empirical distribution of the training data $(\mathbf{x}_1, y_1)\ldots,(\mathbf{x}_n, y_n)$, then we recover the LR problem (2) that has been the primary interest of the previous sections of this paper. On the other hand, it is often useful to conceptualize $\mathcal{D}$ as the *true* underlying distribution of the data $(\mathbf{x}, y)$, in which case $\mathrm{LR}_{\mathcal{D}}$ is the problem of minimizing the expected logistic loss. Both of these abstractions will be useful in our analysis of stochastic gradient descent. We refer to $L_{\mathcal{D}}(\cdot)$ in both cases simply as the "logistic loss function" with the implicit understanding that the function $L_{\mathcal{D}}(\cdot)$ is able to capture both the empirical logistic loss as well as the expected logistic loss. Throughout this section, $H(\cdot)$ denotes the Hessian of $L_{\mathcal{D}}(\cdot)$.

After briefly reviewing stochastic gradient descent (SGD) for smooth convex optimization in Section 4.1, we examine SGD as applied to logistic regression in two cases: in Section 4.2 we consider the non-separable case and examine $\mathrm{LR}_{\mathcal{D}}$ in fully generality, and in Section 4.3 we consider the separable case when $\mathcal{D}$ is the empirical distribution of a training dataset, i.e., the training data problem LR in (2).

4.1 Brief Review of Stochastic Gradient Descent

We first review the stochastic gradient descent method for solving the generic unconstrained differentiable convex optimization problem (14). By way of motivating context, it is sometimes the case that computing the gradient $\nabla f(\cdot)$ of $f(\cdot)$ is very expensive or even intractable, but it may be relatively easy to compute a stochastic estimate of $\nabla f(x)$ at x, which we denote by $\tilde{\nabla} f(x)$, via a stochastic gradient oracle. We say that the oracle computes an unbiased stochastic gradient if

$\mathbb{E}[\tilde{\nabla} f(x) \mid x] = \nabla f(x)$. Notice that by construction $\tilde{\nabla} f(x)$ is a conditional random variable given x. The basic stochastic gradient descent (SGD) method is presented in Algorithm 2, whose structure is the same as steepest descent (Algorithm 1) for the ℓ_2 norm, the only difference being that the stochastic gradient $\tilde{\nabla} f(x^i)$ replaces the exact gradient $\nabla f(x^i)$ in Step (1.). Also, for simplicity, we only consider the case of a constant step-size sequence $\alpha_i := \bar{\alpha} > 0$ for all $i \geq 0$.

Algorithm 2 Stochastic Gradient Descent with constant step-size $\bar{\alpha}$

```
Initialize at $x^0 \in \mathbb{R}^p$, $i \leftarrow 0$, and set the total number of iterations $k \geq 1$.

At iteration $i$:
1. Call stochastic oracle to compute $\tilde{\nabla} f(x^i)$
2. Update:
   $x^{i+1} \leftarrow x^i - \bar{\alpha} \cdot \tilde{\nabla} f(x^i)$

After $k$ iterations:
Option A: Output $\hat{x}^k \leftarrow \frac{1}{k+1} \sum_{i=0}^{k} x^i$.
Option B: Output $\hat{x}^k \leftarrow x^{I_k}$ where $I_k$ is a random variable distributed uniformly on $\{0, 1, \ldots, k\}$.
```

Stochastic gradient descent, and more generally the idea of stochastic approximation, dates back to the seminal work of Robbins and Monro [27]. For recent works related to stochastic gradient descent see, e.g., Nemirovski et al. [20], Bottou [6], Lan [17], Bottou et al. [7], and the references therein. The output of Algorithm 2 using either Option A or Option B depends on the entire sequence $x^0, x^1, \ldots, x^k$. Nevertheless in both options $\hat{x}^k$ can be updated in an online fashion which does not require storage of the entire sequence $x^0, x^1, \ldots, x^k$. Clearly in the case of Option A we have $\hat{x}_k = \frac{k}{k+1}\hat{x}^{k-1} + \frac{1}{k+1}x^k$ for $k \geq 1$, where by convention $\hat{x}^0 = x^0$. In the case of Option B, note that we can construct I_k recursively as follows: given I_{k-1}, which is uniformly distributed on $\{0, \ldots, k-1\}$, we define I_k to be equal to I_{k-1} with probability $\frac{k}{k+1}$ and equal to k with probability $\frac{1}{k+1}$. Then it holds that I_k is uniformly distributed on $\{0, 1, \ldots, k\}$ and is independent of $x^0, x^1, \ldots, x^k$. As pointed out in [12], Option B can also be implemented by first generating I_k uniformly at random on $\{0, \ldots, k\}$ during the initialization stage of the algorithm, and then only running for I_k iterations before stopping (assuming k is fixed in advance).

In addition to the smoothness condition (15) (with respect to the ℓ_2 norm in this case), a condition that is required in the typical analysis of SGD is the following bounded second moment condition:

$$\mathbb{E}\left[\|\tilde{\nabla} f(x)\|_2^2 \mid x\right] \leq M^2 \quad \text{for all } x \in \mathbb{R}^p \ , \tag{22}$$

where M is a positive constant. In the case of smooth convex optimization, as is studied in [17] for example, (22) is often replaced with a bound on the variance of the stochastic gradient oracle instead, which is smaller. However for our purposes in studying logistic regression, (22) is adequate and simplifies the analysis.

The following theorem presents a stylized version of computational guarantees for SGD, which is an amended and extended version of mostly well-known results about stochastic gradient descent, see for example Nemirovski et al. [20] as well as Ghadimi and Lan [12]. In the theorem, the expectation in items *(i)* and *(iii)* is taken with respect to all of the stochasticity of Algorithm 2.

Theorem 4.1. (Computational Guarantees for Stochastic Gradient Descent) *Let $\{x^k\}$ be generated according to the stochastic gradient descent method (Algorithm 2) using a constant step-size $\bar{\alpha} > 0$. Under either Option A or Option B, it holds for all $k \geq 0$ that:*

(i) (expected optimality gap):

$$\mathbb{E}[f(\hat{x}^k)] - f(x) \ \leq \ \frac{\|x^0 - x\|_2^2}{2\bar{\alpha}(k+1)} + \frac{\bar{\alpha}M^2}{2} \quad \textit{for all } x \in \mathbb{R}^p \ , \ \textit{and}$$

(ii) (norm bound):

$$\|x^k - x^0\|_2 \ \leq \ \bar{\alpha}\sum_{i=0}^{k-1} \|\tilde{\nabla} f(x^i)\|_2 \ .$$

Under Option B, it holds for all $k \geq 0$ that:

(iii) (expected gradient bound):

$$\mathbb{E}\left[\|\nabla f(\hat{x}^k)\|_2^2\right] \ \leq \ \frac{f(x^0) - f^*}{\bar{\alpha}(k+1)} + \frac{\bar{\alpha}LM^2}{2} \ .$$

□

For completeness, a self-contained proof of Theorem 4.1 is given in Appendix A.6. Note that when an optimal solution x^* of (14) exists, then we can take $x \leftarrow x^*$ in item *(i)* to obtain a bound on the expected optimality gap $\mathbb{E}[f(\hat{x}^k)] - f^*$. Items *(i)* and *(iii)* of Theorem 4.1 present bounds that hold in expectation; bounds that hold with high probability require additional assumptions such as moment generating function type assumptions, compactness of the feasible region, etc., see [20] and [12].

4.2 Informing Stochastic Gradient Descent for Logistic Regression in the Non-Separable Case

Let us now return to the logistic regression problem $\mathrm{LR}_{\mathcal{D}}$, which we examine in full generality in the case of non-separable data. First we need to extend the definitions of non-separable and separable datasets to an arbitrary distribution $\mathcal{D}$ over the data $(\mathbf{x}, y)$. Let $\mathrm{supp}(\mathcal{D}) \subseteq \mathbb{R}^p \times \{-1, +1\}$ denote the support of the distribution $\mathcal{D}$. Then we say that the data distribution $\mathcal{D}$ is *separable* if there exists a model $\beta \in \mathbb{R}^p$ such that $\inf_{(\mathbf{x},y)\in\mathrm{supp}(\mathcal{D})} y\beta^T\mathbf{x} > 0$. Otherwise, if $\inf_{(\mathbf{x},y)\in\mathrm{supp}(\mathcal{D})} y\beta^T\mathbf{x} \leq 0$ for every model β, then we say that the data distribution $\mathcal{D}$ is *non-separable*.

As in the previously examined case of finite training datasets, clearly some non-separable distributions might be "more non-separable" than others, so let us introduce a way to measure the extent to which the distribution is non-separable. We define the degree of non-separability of the distribution $\mathcal{D}$ (with respect to the norm $\|\cdot\|$) to be:

$$\begin{array}{rll} \mathrm{DegNSEP}^*_{\mathcal{D}} := & \displaystyle\min_{\beta\in\mathbb{R}^p} & \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[[y\beta^T\mathbf{x}]^-\right] \\ & \text{s.t.} & \|\beta\| = 1 \ , \end{array} \tag{23}$$

which states that $\mathrm{DegNSEP}^*_{\mathcal{D}}$ is the smallest (over all normalized models β) expected misclassification error of the model β. Note that the norm $\|\cdot\|$ in the above definition is any generic given norm. It is straightforward to extend Proposition 2.1 to this more general setting, and we present this generalization in Proposition 4.1.

Proposition 4.1. *If* $\mathrm{DegNSEP}^*_{\mathcal{D}} > 0$*, then:*

(i) there is a unique optimal solution β^* *of the logistic regression problem* $\mathrm{LR}_{\mathcal{D}}$*,*

(ii) $H(\beta^*) \succ 0$,

(iii) $\|\beta^*\| \ \le \ \dfrac{L^*_{\mathcal{D}}}{\mathrm{DegNSEP}^*_{\mathcal{D}}} \ \le \ \dfrac{\ln(2)}{\mathrm{DegNSEP}^*_{\mathcal{D}}}$ *, and*

(iv) $\mathrm{Dist}_0 \ \le \ \dfrac{\ln(2) + L^*_{\mathcal{D}}}{\mathrm{DegNSEP}^*_{\mathcal{D}}} \ \le \ \dfrac{2\ln(2)}{\mathrm{DegNSEP}^*_{\mathcal{D}}}$.

Proof: The proof follows the same structure as the proof of Proposition 2.1, but requires a more careful argument to prove that $H(\beta) \succ 0$ for all $\beta \in \mathbb{R}^p$. Suppose that $\mathrm{DegNSEP}^*_{\mathcal{D}} > 0$, and let $\beta \in \mathbb{R}^p$ be given. It can also be demonstrated (e.g., by Section 7.2.4 of [29]) that $L_{\mathcal{D}}(\cdot)$ is twice differentiable and that $H(\beta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell''(y\beta^T\mathbf{x})\mathbf{x}\mathbf{x}^T\right]$. Note that $H(\beta) \succeq 0$ by convexity. Suppose, by way of contradiction, that there exists $\bar{\beta} \in \mathbb{R}^p$ with $\|\bar{\beta}\| = 1$ and $\bar{\beta}^T H(\beta)\bar{\beta} = 0$. Then we have that $0 = \bar{\beta}^T H(\beta)\bar{\beta} = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell''(y\beta^T\mathbf{x})\bar{\beta}^T\mathbf{x}\mathbf{x}^T\bar{\beta}\right] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell''(y\beta^T\mathbf{x})(\bar{\beta}^T\mathbf{x})^2\right]$. Therefore it holds that $\ell''(y\beta^T\mathbf{x})(\bar{\beta}^T\mathbf{x})^2 = 0$ with probability one, and since $\ell''(y\beta^T\mathbf{x}) > 0$ we must have that $\bar{\beta}^T\mathbf{x} = 0$ with probability one. This then implies that $\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[[y\bar{\beta}^T\mathbf{x}]^-\right] = 0$, which implies that $\mathrm{DegNSEP}^*_{\mathcal{D}} = 0$, and this provides the desired contradiction. Therefore $H(\beta) \succ 0$. The remainder of the proof exactly follows that of Proposition 2.1, and is omitted for brevity. □

Throughout this section we assume the following finiteness condition on second moments of $\mathcal{D}$:

Assumption 4.1. (Finite second moments of the data distribution $\mathcal{D}$) *The data distribution* $\mathcal{D}$ *satisfies* $\mathbb{E}\left[\|\mathbf{x}\|_2^2\right] < +\infty$.

Define the second moment matrix $\Sigma := \mathbb{E}[\mathbf{x}\mathbf{x}^T] \in \mathbb{R}^{p\times p}$. It follows from Assumption 4.1 that Σ is well-defined and finite, and that $L_{\mathcal{D}}(\cdot)$ is continuous, convex, differentiable, and satisfies $\nabla L_{\mathcal{D}}(\beta) = \mathbb{E}\left[\nabla_\beta \ln\left(1 + \exp\left(-y\beta^T\mathbf{x}\right)\right)\right]$ for all $\beta \in \mathbb{R}^p$ (see, e.g., [29]). We also have:

Proposition 4.2. (Lipschitz smoothness of the logistic loss function) *The logistic loss function* $L_{\mathcal{D}}(\cdot)$ *is* $L = \frac{1}{4}\lambda_{\max}(\Sigma)$*-smooth with respect to the* ℓ_2 *norm on* $\mathbb{R}^p$*.* □

A proof of Proposition 4.2 is given in Appendix A.7.

Denote the scalar logistic loss function by $\ell(t) := \ln(1+\exp(-t))$. We assume the following regarding the stochastic gradient oracle for the logistic loss function $L_{\mathcal{D}}(\beta)$ of (21):

Assumption 4.2. (Stochastic gradient oracle for the logistic loss function) *The stochastic gradient oracle* $\tilde{\nabla} L_{\mathcal{D}}(\cdot)$ *is implemented by drawing an independent sample from the distribution* $\mathcal{D}$*. That is, for any (possibly random)* $\beta \in \mathbb{R}^p$*, the stochastic gradient* $\tilde{\nabla} L_{\mathcal{D}}(\beta)$ *is computed by independently sampling* $(\tilde{\mathbf{x}}, \tilde{y})$ *from the distribution* $\mathcal{D}$ *and assigning* $\tilde{\nabla} L_{\mathcal{D}}(\beta) \leftarrow \nabla_\beta \ell(\tilde{y}\beta^T\tilde{\mathbf{x}})$*.*

Proposition 4.3. (Second moment of the stochastic gradient) *The stochastic gradient*

$\tilde{\nabla} L_{\mathcal{D}}(\cdot)$ *computed from the oracle described in Assumption 4.2 satisfies the second moment upper bound* (22) *with* $M^2 = \mathrm{Tr}(\Sigma)$. □

A proof of Proposition 4.3 is given in Appendix A.8.

Theorem 4.2 presents the main computational guarantees for SGD applied to the logistic regression problem (21) in the case when the data distribution $\mathcal{D}$ is non-separable.

Theorem 4.2. (Computational Guarantees for SGD: Non-Separable Case) *Suppose that SGD (Algorithm 2) for logistic regression is initialized at* $\beta^0 := 0$ *and is implemented using the constant step-size* $\bar{\alpha} > 0$, *and that the stochastic gradients are computed as in Assumption 4.2. Under either Option A or Option B of SGD (Algorithm 2), if the data distribution* $\mathcal{D}$ *is non-separable, then for all* $k \geq 0$ *it holds that:*

$$\mathbb{E}\left[L_{\mathcal{D}}(\hat{\beta}^k)\right] - L^*_{\mathcal{D}} \ \leq \ \frac{(\ln(2))^2}{2\bar{\alpha}(k+1)\cdot(\mathrm{DegNSEP}^*_{\mathcal{D}})^2} \ + \ \frac{\bar{\alpha}\cdot\mathrm{Tr}(\Sigma)}{2} \ . \tag{24}$$

Proof: This result is a straightforward application of Theorem 4.1, Proposition 4.1, and Proposition 4.3. By item *(i)* of Proposition 4.1 an optimal solution β^* exists. The result follows by directly applying item *(i)* of Theorem 4.1, along with the bound $\|\beta^*\| \leq \frac{\ln(2)}{\mathrm{DegNSEP}^*_{\mathcal{D}}}$ provided by item *(ii)* of Proposition 4.1, and the value of $M^2 = \mathrm{Tr}(\Sigma)$ from Proposition 4.3. □

Theorem 4.2 provides an upper bound on the expected logistic loss that naturally depends on the condition number $\mathrm{DegNSEP}^*_{\mathcal{D}}$ and that holds for any step-size value $\bar{\alpha}$. Given knowledge of the constants $\mathrm{DegNSEP}^*_{\mathcal{D}}$ and $\mathrm{Tr}(\Sigma)$, it is possible to tune $\bar{\alpha}$ in order to minimize the upper bound expression on the right side of (24) for a given k. However, in practice one does not typically know either of these constants. Indeed, it is more realistic to assume one has knowledge of a deterministic upper bound on the size of the feature vectors, i.e., there is an available constant $R > 0$ for which $\|\mathbf{x}\|_2 \leq R$ with probability one. Under this assumption it also follows that $\mathrm{Tr}(\Sigma) = \mathbb{E}[\|\mathbf{x}\|_2^2] \leq R^2$. Corollary 4.1 presents a computational guarantee for SGD in the case of non-separable data under a slightly weaker assumption and using a step-size that only incorporates knowledge of the constant R.

Corollary 4.1. *Suppose that we have available a constant* R *such that* $\mathrm{Tr}(\Sigma) \leq R^2$. *Consider running SGD (Algorithm 2) for a total of* k *iterations, initialized at* $\beta^0 := 0$, *with stochastic gradients computed as in Assumption 4.2, and using the constant step-size*

$$\bar{\alpha} := \frac{\ln(2)}{R^2\sqrt{k+1}} \ .$$

Under either Option A or Option B of SGD (Algorithm 2), if the data distribution $\mathcal{D}$ *is non-separable it holds for all* $k \geq 0$ *that:*

$$\mathbb{E}\left[L_{\mathcal{D}}(\hat{\beta}^k)\right] - L^*_{\mathcal{D}} \ \leq \ \frac{\ln(2)}{2\sqrt{k+1}}\left(\frac{R^2}{(\mathrm{DegNSEP}^*_{\mathcal{D}})^2} + 1\right) \ .$$

□

Theorem 4.2 and Corollary 4.1 present results that highlight the role of the condition number $\mathrm{DegNSEP}^*_{\mathcal{D}}$ in the well-known $O(1/\sqrt{k})$ computational guarantee for stochastic gradient descent.

It is also possible to study how $\mathrm{DegNSEP}^*_{\mathcal{D}}$ informs the adaptive $O(1/\mu k)$ (where μ is the smallest eigenvalue of the Hessian at the optimum) guarantees developed by Bach in [3]. Proposition 4.4 below directly follows from Proposition 10 of [3] and demonstrates how $\mathrm{DegNSEP}^*_{\mathcal{D}}$ informs the corresponding computational guarantees developed therein. Note that $\lambda_{\min}(H(\beta^*)) > 0$ by part *(ii)* of Proposition 4.1. Note that the step-size considered in Proposition 4.4 is larger than that considered in [3] by a constant factor of $2\ln(2)$.

Proposition 4.4. (Bach [3], Proposition 10) *Suppose that we have available a constant R such that $\|\mathbf{x}\|_2 \le R$ with probability one. Consider running SGD (Algorithm 2) for a total of k iterations, initialized at $\beta^0 := 0$, with stochastic gradients computed as in Assumption 4.2, and using the constant step-size*

$$\bar{\alpha} := \frac{\ln(2)}{R^2\sqrt{k+1}} \ .$$

Under Option A of SGD (Algorithm 2), if the data distribution $\mathcal{D}$ is non-separable and $\mathrm{DegNSEP}^*_{\mathcal{D}} > 0$, then it holds for all $k \ge 0$ that:

$$(i) \qquad \mathbb{E}\left[L_{\mathcal{D}}(\hat{\beta}^k)\right] - L^*_{\mathcal{D}} \ \le \ \frac{R^2}{\lambda_{\min}(H(\beta^*))(k+1)}\left(\frac{10R(\ln(2))^2}{\mathrm{DegNSEP}^*_{\mathcal{D}}} + 15\right)^4 \ , \textit{ and}$$

$$(ii) \qquad \mathbb{E}\left[\|\hat{\beta}^k - \beta^*\|_2^2\right] \ \le \ \frac{R^2}{\lambda_{\min}(H(\beta^*))^2(k+1)}\left(\frac{12R(\ln(2))^2}{\mathrm{DegNSEP}^*_{\mathcal{D}}} + 21\right)^4 \ ,$$

where β^* is the unique optimal solution of $\mathrm{LR}_{\mathcal{D}}$.

Proof: Consider the scaled objective function $\tilde{L}_{\mathcal{D}}(\cdot) := 2\ln(2)L_{\mathcal{D}}(\cdot)$, and denote its Hessian by $\tilde{H}(\cdot)$. It can be verified that $\tilde{L}_{\mathcal{D}}(\cdot)$ satisfies assumptions (A1) - (A7) of [3] with constant $2\ln(2)R$. Clearly, running SGD with objective $L_{\mathcal{D}}(\cdot)$ and constant step-size $\frac{\ln(2)}{R^2\sqrt{k+1}}$ is equivalent to running SGD with objective $\tilde{L}_{\mathcal{D}}(\cdot)$ and constant step-size $\frac{1}{2R^2\sqrt{k+1}}$, which is required by Proposition 10 of [3]. Therefore, directly applying this proposition to $\tilde{L}_n(\cdot)$ yields:

$$\mathbb{E}\left[\tilde{L}_{\mathcal{D}}(\hat{\beta}^k)\right] - \tilde{L}^*_{\mathcal{D}} \ \le \ \frac{(2\ln(2))^2R^2}{\lambda_{\min}(\tilde{H}(\beta^*))(k+1)}\left(10R\ln(2)\|\beta^*\|_2 + 15\right)^4 \ .$$

Finally, using $\tilde{H}(\beta^*) = 2\ln(2)H(\beta^*)$, the upper bound $\|\beta^*\| \le \frac{\ln(2)}{\mathrm{DegNSEP}^*_{\mathcal{D}}}$ provided by item *(ii)* of Proposition 4.1, and dividing both sides of the above by $2\ln(2)$ yields part *(i)* of the theorem. Part *(ii)* also directly follows from Proposition 10 of [3] by a similar argument. □

Comparing Corollary 4.1 with Proposition 4.4, note that the relative sizes of the constants R, $\mathrm{DegNSEP}^*_{\mathcal{D}}$, and $\lambda_{\min}(H(\beta^*))$, as well as the total number of iterations k will determine which bound dominates the other. And of course if k is large enough then Proposition 4.4 yields the better bound.

4.3 Informing Stochastic Gradient Descent for Logistic Regression in the Separable Case

In this subsection we examine the properties of SGD in the case when the data is separable, for the previously examined logistic regression problem (2), i.e., we assume that $\mathcal{D}$ is the empirical distribution of the training data $(\mathbf{x}_1, y_1)\dots,(\mathbf{x}_n, y_n)$ in this subsection and we revert to the notation used throughout Section 3. The following theorem presents our results in this case.

Theorem 4.3. (Computational Guarantees for SGD: Separable Case) *Suppose that we have available a constant R such that $\|\mathbf{x}_i\|_2 \le R$ for all $i \in \{1,\dots,n\}$. Consider running SGD (Algorithm 2) for a total of k iterations, initialized at $\beta^0 := 0$, with stochastic gradients computed as in Assumption 4.2, and using the constant step-size*

$$\bar{\alpha} := \frac{\ln(2)}{R^2\sqrt{k+1}} \ .$$

Under Option B of SGD (Algorithm 2), when the data is separable it holds for all $k \ge 1$ that:

(i) (margin bound): For any $\gamma \in (0,1]$, with probability at least $1-\gamma$ the normalized iterate $\bar{\beta}^k := \hat{\beta}^k/\|\hat{\beta}^k\|$ satisfies

$$\rho(\bar{\beta}^k) \ > \ \frac{\text{DegSEP}^* \cdot \ln\left(\frac{\text{DegSEP}^* \sqrt{\gamma}\,\sqrt[4]{k+1}}{nR\sqrt{1.1}} - 1\right)}{2(\ln(k)+1)} \tag{25}$$

(ii) (shrinkage): $\|\hat{\beta}^k\|_2 \ \le \ \dfrac{2\ln(k)}{\text{DegSEP}^*} + \dfrac{2}{\|\mathbf{X}\|_{2,\infty}}$, and

(iii) (expected gradient bound): $\mathbb{E}\left[\|\nabla L_n(\hat{\beta}^k)\|_2^2\right] \ < \ \dfrac{1.1 \cdot R^2}{\sqrt{k+1}}$. □

The margin bound in Theorem 4.3 is similar in flavor to the bound for steepest descent in Theorem 3.4, but is weaker (due to stochasticity). Indeed, by similar arguments as in Section 3.4, the bound in item *(i)* of Theorem 4.3 implies a computational guarantee of the form

$$\frac{\rho(\bar{\beta}^k)}{\frac{\text{DegSEP}^*}{8}} \ge 1 - \frac{C}{\ln(k+1)} \ \ \text{with probability at least } 1-\gamma \ , \tag{26}$$

for any fixed $\gamma \in (0,1]$, with $C = 4\ln(2) - 2\ln(\gamma) + \ln(\frac{(1.1)^2 n^4 R^4}{(\text{DegSEP}^*)^4}) + 1$. (This should be compared to the case of deterministic steepest descent where we have deterministic convergence to at least $\frac{\text{DegSEP}^*}{4}$.) To demonstrate this, consider setting $k := \lfloor \frac{\Omega^4 (1.1)^2 n^4 R^4}{(\text{DegSEP}^*)^4 \gamma^2} \rfloor$ for some parameter $\Omega \ge 2$. Then the bound in (25) becomes:

$$\rho(\bar{\beta}^k) \ \ge \ \frac{\text{DegSEP}^* \cdot \ln(\Omega - 1)}{8\ln(\Omega) - 4\ln(\gamma) + 2\ln\left(\frac{(1.1)^2 n^4 R^4}{(\text{DegSEP}^*)^4}\right) + 2} \ ,$$

and rearranging the above and using $\Omega \ge 2$ yields:

$$\begin{aligned}
\frac{\rho(\bar{\beta}^k)}{\frac{\text{DegSEP}^*}{8}} \ &\ge \ 1 \ - \ \frac{\ln(2) - \frac{1}{2}\ln(\gamma) + \frac{1}{4}\ln\left(\frac{(1.1)^2 n^4 R^4}{(\text{DegSEP}^*)^4}\right) + \frac{1}{4}}{\ln(\Omega) - \frac{1}{2}\ln(\gamma) + \frac{1}{4}\ln\left(\frac{(1.1)^2 n^4 R^4}{(\text{DegSEP}^*)^4}\right) + \frac{1}{4}} \\
&\ge \ 1 \ - \ \frac{\ln(2) - \frac{1}{2}\ln(\gamma) + \frac{1}{4}\ln\left(\frac{(1.1)^2 n^4 R^4}{(\text{DegSEP}^*)^4}\right) + \frac{1}{4}}{\frac{1}{4}\ln(k+1)} \ = \ 1 - \frac{C}{\ln(k+1)} \ .
\end{aligned}$$

This result is similar to results in [19], wherein $O(1/\ln(k))$ convergence towards DegSEP^* is demonstrated for SGD for sampling without replacement. Note that we provide an explicit margin bound in item *(i)* above and that we study SGD with sampling with replacement. On the other hand, [19] of course proves convergence towards DegSEP^*, not $\frac{\text{DegSEP}^*}{8}$. It would be interesting to see if the tools used to prove all of these results can be combined somehow to yield stronger convergence guarantees about the margin for SGD. Note also that the margin bound (25) is proven by applying Markov's inequality with the bound on the second moment of $\|\nabla L_n(\hat{\beta}^k)\|_2$ given by item *(iii)* of the theorem. In the case of non-separable data, Bach [3] is able to strengthen this second moment bound to $O(1/k)$ and also derives bounds on the higher-order moments of $\|\nabla L_n(\hat{\beta}^k)\|_2$ (for Option A of SGD). It would also be interesting to see if similar bounds can be derived and used to strengthen the margin bound in the case of separable data.

In order to prove Theorem 4.3, we will use the following lemma, which is similar to Lemma 3.1 and bounds the norms of iterates of SGD applied to the logistic regression problem (2).

Lemma 4.1. (essentially from Ji and Telgarsky [16]) *Suppose that SGD (Algorithm 2) is initialized at* $\beta^0 := 0$ *using the constant step-size value* $\bar{\alpha}$. *If* $\text{DegSEP}^* > 0$ *and* $\bar{\alpha} \leq \frac{2}{\|\mathbf{X}\|_{2,\infty}^2}$, *then it holds for all* $k \geq 1$ *that:*

$$\|\beta^k\|_2 \ \leq \ \frac{2\ln(k)}{\text{DegSEP}^*} + \frac{2}{\|\mathbf{X}\|_{2,\infty}} \ .$$

□

The proof of this lemma is presented in Appendix A.5.

Proof of Theorem 4.3: Item *(ii)* follows directly from Lemma 4.1 as well as the fact that $\ln(\cdot)$ is an increasing function. Item *(iii)* is a straightforward application of item *(iii)* of Theorem 4.1. Indeed, recalling that $L_n(\beta^0) - L_n^* = \ln(2)$ (equality holds in the separable case) and $\lambda_{\max}(\Sigma) \leq \text{Tr}(\Sigma) \leq R^2$, item *(iii)* follows directly from the definition of $\bar{\alpha}$, Propositions 4.2 and 4.3, and item *(iii)* of Theorem 4.1.

To prove item *(i)* first note that Markov's inequality yields:

$$\mathbb{P}\left(\|\nabla L_n(\hat{\beta}^k)\|_2^2 \geq \frac{1.1 \cdot R^2}{\gamma\sqrt{k+1}}\right) \leq \frac{\mathbb{E}\left[\|\nabla L_n(\hat{\beta}^k)\|_2^2\right]}{\frac{1.1 \cdot R^2}{\gamma\sqrt{k+1}}} < \gamma \ ,$$

where the second inequality follows from item *(iii)* of the theorem. Therefore with probability at least $1 - \gamma$ it holds that:

$$\|\nabla L_n(\hat{\beta}^k)\|_2^2 < \frac{1.1 \cdot R^2}{\gamma\sqrt{k+1}} \ . \tag{27}$$

We will now demonstrate that if (27) holds (in addition to everything else) then (25) holds, which therefore implies that the statement in part *(i)* is true. Indeed, combining (27) with Lemma 2.1 yields:

$$\rho(\hat{\beta}^k) \ \geq \ \ln\left(\frac{\text{DegSEP}^*}{n\|\nabla L_n(\hat{\beta}^k)\|_2} - 1\right) \ > \ \ln\left(\frac{\text{DegSEP}^*\sqrt{\gamma}\sqrt[4]{k+1}}{nR\sqrt{1.1}} - 1\right) \ .$$

Combining the above with item *(ii)* and using $\text{DegSEP}^* \leq \|\mathbf{X}\|_{2,\infty}$, we obtain:

$$\rho(\bar{\beta}^k) = \frac{\rho(\hat{\beta}^k)}{\|\hat{\beta}^k\|_2} \ > \ \frac{\ln\left(\frac{\text{DegSEP}^*\sqrt{\gamma}\sqrt[4]{k+1}}{nR\sqrt{1.1}} - 1\right)}{\frac{2\ln(k)}{\text{DegSEP}^*} + \frac{2}{\|\mathbf{X}\|_{2,\infty}}} \ \geq \ \frac{\ln\left(\frac{\text{DegSEP}^*\sqrt{\gamma}\sqrt[4]{k+1}}{nR\sqrt{1.1}} - 1\right)}{\frac{2\ln(k)}{\text{DegSEP}^*} + \frac{2}{\text{DegSEP}^*}} \ ,$$

and the proof then follows from rearranging terms in the above inequality. □

5 Conclusions

The theme of this paper is the interplay between data conditioning, behavior/properties of the optimization problem, and computational guarantees of first-order methods, all in the context of logistic regression. We have presented results that make rigorous the intuitive notion that the optimization problem itself as well as the corresponding algorithms for training a logistic regression model are well-behaved when the degree of non-separability of the dataset is large. We also have presented results that demonstrate that the specific algorithmic properties of steepest descent and stochastic gradient descent lead to large margin solutions in the case of separable data, which runs counter to the intuition that logistic regression is ill-behaved in this case. We hope that further examination of the role of data and problem conditioning in the analysis of other statistical learning problems and other algorithms will extend the general understanding of these problems and algorithms.

References

[1] A. Albert and J. A. Anderson. On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71(1):1–10, 1984.

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

[3] F. Bach. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *The Journal of Machine Learning Research*, 15(1):595–627, 2014.

[4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

[5] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM Journal on Optimization*, 23(2):2037—-2060, 2013.

[6] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[7] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[8] E. J. Candès and P. Sur. The phase transition for the existence of the maximum likelihood estimate in high-dimensional logistic regression. *arXiv preprint arXiv:1804.09753*, 2018.

[9] T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE transactions on electronic computers*, (3):326–334, 1965.

[10] R. Freund and P. Grigas. New analysis and results for the Frank-Wolfe method. *Mathematical Programming*, 155:199–230, 2016.

[11] J. Friedman, T. Hastie, and R. Tibshirani. Special invited paper. additive logistic regression: A statistical view of boosting. *The Annals of Statistics*, 28(2):337–374, 2000.

[12] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[13] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[14] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Verlag, New York, 2009.

[16] Z. Ji and M. Telgarsky. Risk and parameter convergence of logistic regression. Technical report, University of Illinois, Urbana-Champlain, Urbana-Champlain, IL, 2018.

[17] G. Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1-2):365–397, 2012.

[18] M. S. Nacson, J. Lee, S. Gunasekar, P. H. P. Savarese, N. Srebro, and D. Soudry. Convergence of gradient descent on separable data. Technical report, arXiv:1803.01905v2, 2018.

[19] M. S. Nacson, N. Srebro, and D. Soudry. Stochastic gradient descent on separable data: Exact convergence with a fixed learning rate. Technical report, arXiv:1806.01796v1, 2018.

[20] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[21] Y. Nesterov. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM Journal on Optimization*, 22(2):341–362, 2012.

[22] Y. E. Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied Optimization*. Kluwer Academic Publishers, Boston, 2003.

[23] Y. E. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

[24] B. Polyak. *Introduction to Optimization*. Optimization Software, Inc., New York, 1987.

[25] J. Renegar. Some perturbation theory for linear programming. *Mathematical Programming*, 65(1):73–91, 1994.

[26] P. Richtárik and M. Takáč. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1):1–38, 2014.

[27] H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Statist.*, 22(3):400–407, 09 1951.

[28] N. L. Roux, M. Schmidt, and F. Bach. Stochastic gradient method with an exponential convergence rate for finite training sets. In *Advances in Neural Information Processing Systems*. 2012.

[29] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[30] M. J. Silvapulle. On the existence of maximum likelihood estimators for the binomial response models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 310–313, 1981.

[31] D. Soudry, E. Hoffer, and N. Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

[32] J. Zhu and T. Hastie. Kernel logistic regression and the import vector machine. *Journal of Computational and Graphical Statistics*, 14(1):185–205, 2005.

A Appendix

A.1 Proofs of Propositions 2.4 and 2.5

Proof of Proposition 2.4: Notice that the optimization problem defining (4) has a continuous objective function and a compact feasible region, whereby it follows from the Weierstrass Theorem that (4) attains its optimum at some $\bar{\beta}$ and therefore $\mathrm{DegNSEP}^* = \frac{1}{n}\sum_{i=1}^{n}[y_i \mathbf{x}_i^T \bar{\beta}]^-$. It follows from norm duality that there exists $\bar{s}$ satisfying $\|\bar{s}\|_* = 1$ and $\bar{s}^T\bar{\beta} = \|\bar{\beta}\| = 1$. Let $\varepsilon > 0$ be given, and now define $\Delta\mathbf{X} := u\bar{s}^T$ where $u_i := y_i[y_i\mathbf{x}_i^T\bar{\beta}]^- + y_i\varepsilon$. Notice for each $i = 1, \ldots, n$ that $y_i(\mathbf{x}_i + \Delta\mathbf{x}_i)^T\bar{\beta} = y_i\mathbf{x}_i^T\bar{\beta} + y_i u_i \bar{s}^T\bar{\beta} = y_i\mathbf{x}_i^T\bar{\beta} + [y_i\mathbf{x}_i^T\bar{\beta}]^- + \varepsilon \geq \varepsilon > 0$, whereby the perturbed dataset $(\mathbf{X}+\Delta\mathbf{X}, y)$ is separable and hence $\mathrm{PertSEP}^* \leq \frac{1}{n}\|\Delta\mathbf{X}\|_{\cdot,1} = \frac{1}{n}\|u\|_1\|\bar{s}\|_* = \frac{1}{n}\|u\|_1 = \mathrm{DegNSEP}^* + \varepsilon$. As this is true for any $\varepsilon > 0$ it follows that $\mathrm{PertSEP}^* \leq \mathrm{DegNSEP}^*$.

We next show that $\mathrm{DegNSEP}^* \leq \mathrm{PertSEP}^*$, which will complete the proof. Suppose that $\Delta\mathbf{X}$ satisfies $(\mathbf{X} + \Delta\mathbf{X}, y)$ is separable, and hence there exists β with $\|\beta\| = 1$ and $y_i(\mathbf{x}_i + \Delta\mathbf{x}_i)^T\beta > 0$ for $i = 1, \ldots, n$. Define the vector v component-wise for $i = 1, \ldots, n$ by:

$$v_i := \begin{cases} 0 & \text{if } \ y_i\mathbf{x}_i^T\beta \geq 0 \\ y_i & \text{if } \ y_i\mathbf{x}_i^T\beta < 0 \ , \end{cases}$$

and notice in particular that if $y_i\mathbf{x}_i^T\beta < 0$ then $[y_i\mathbf{x}_i^T\beta]^- = -y_i\mathbf{x}_i^T\beta < y_i(\Delta\mathbf{x}_i)^T\beta = v_i(\Delta\mathbf{x}_i)^T\beta$. Also, if $y_i\mathbf{x}_i^T\beta \geq 0$, then $[y_i\mathbf{x}_i^T\beta]^- = 0 = v_i(\Delta\mathbf{x}_i)^T\beta$. Therefore $\mathrm{DegNSEP}^* \leq \frac{1}{n}\sum_{i=1}^{n}[y_i\mathbf{x}_i^T\beta]^- \leq \frac{1}{n}\sum_{i=1}^{n} v_i(\Delta\mathbf{x}_i)^T\beta = \frac{1}{n}v^T\Delta\mathbf{X}\beta \leq \frac{1}{n}\|\Delta\mathbf{X}\|_{\cdot,1}$ since $\|v\|_\infty \leq 1$. Thus $\mathrm{DegNSEP}^* \leq \frac{1}{n}\|\Delta\mathbf{X}\|_{\cdot,1}$ for any perturbation $\Delta\mathbf{X}$ for which $(\mathbf{X} + \Delta\mathbf{X}, y)$ is separable, and hence $\mathrm{DegNSEP}^* \leq \mathrm{PertSEP}^*$, completing the proof. □

Proof of Proposition 2.5: Define $\Delta_n = \{\lambda \in \mathbb{R}^n : e^T\lambda = 1, \ \lambda \geq 0\}$. We can write (8) in maxmin form as:

$$\mathrm{DegSEP}^* := \max_{\beta:\|\beta\|\leq 1} \ \min_{\lambda\in\Delta_n} \lambda^T Y\mathbf{X}\beta \ \ = \ \ \min_{\lambda\in\Delta_n} \ \max_{\beta:\|\beta\|\leq 1} \ \lambda^T Y\mathbf{X}\beta \ \ = \ \ \min_{\lambda\in\Delta_n} \|\mathbf{X}^T Y\lambda\|_* \ , \qquad (28)$$

where the middle equality follows from minmax strong duality. Furthermore, both the minmax problem and the maxmin problem attain their optima for some $\bar{\beta}$ satisfying $\|\bar{\beta}\| \leq 1$ and $\bar{\lambda} \in \Delta_n$ which implies that:

$$\text{DegSEP}^* \ := \ \bar{\lambda}^T Y \mathbf{X} \bar{\beta} \ = \ \rho(\bar{\beta}) \ = \min_i (y_i \mathbf{x}_i^T \bar{\beta})_i \ = \ \|\mathbf{X}^T Y \bar{\lambda}\|_* \ . \tag{29}$$

Now define $\Delta\mathbf{X} := -y\bar{\lambda}^T Y \mathbf{X}$. Direct substitution yields $\bar{\lambda}^T Y(\mathbf{X} + \Delta\mathbf{X}) = 0$, which then implies that there does not exist any β satisfying $Y(\mathbf{X}+\Delta\mathbf{X})\beta > 0$, and hence $(\mathbf{X}+\Delta\mathbf{X}, y)$ is not separable. Therefore $\text{PertNSEP}^* \leq \|\Delta\mathbf{X}\|_{\cdot,\infty} = \|-y\bar{\lambda}^T Y \mathbf{X}\|_{\cdot,\infty} = \|y\|_\infty \|\mathbf{X}^T Y \bar{\lambda}\|_* = \text{DegSEP}^*$.

We next show that $\text{DegSEP}^* \leq \text{PertNSEP}^*$, which will complete the proof. Suppose that $\Delta\mathbf{X}$ satisfies $(\mathbf{X} + \Delta\mathbf{X}, y)$ is not separable, and hence by a theorem of the alternative there exists $\lambda \in \Delta_n$ satisfying $\lambda^T Y(\mathbf{X} + \Delta\mathbf{X}) = 0$. Using the values $\bar{\beta}$ and $\bar{\lambda}$ defined above, we have:

$$\text{DegSEP}^* = \bar{\lambda}^T Y \mathbf{X} \bar{\beta} \leq \lambda^T Y \mathbf{X} \bar{\beta} = -\lambda^T Y \Delta\mathbf{X} \bar{\beta} \leq \|\Delta\mathbf{X}\|_{\cdot,\infty} \|\bar{\beta}\| \|Y\lambda\|_1 \leq \|\Delta\mathbf{X}\|_{\cdot,\infty} \ .$$

Thus $\text{DegSEP}^* \leq \|\Delta\mathbf{X}\|_{\cdot,\infty}$ for any perturbation $\Delta\mathbf{X}$ for which $(\mathbf{X} + \Delta\mathbf{X}, y)$ is non-separable, and hence $\text{DegSEP}^* \leq \text{PertNSEP}^*$, completing the proof. □

A.2 Proof of Theorem 3.1

Since $f(\cdot)$ satisfies (15), it follows easily from the fundamental theorem of calculus that:

$$f(y) \leq f(x) + \nabla f(x)^T (y - x) + \tfrac{L}{2}\|y - x\|^2 \quad \text{for all } x, y \ . \tag{30}$$

(For a short proof of this fact, see Proposition A.2 of [10] for example.) Applying (30) to the iterates of the Steepest Descent Method yields the following for each $i \geq 0$:

$$\begin{aligned} f(x^{i+1}) &\leq f(x^i) + \nabla f(x^i)^T (x^{i+1} - x^i) + \tfrac{L}{2}\|x^{i+1} - x^i\|^2 \\ &= f(x^i) - \alpha_i \|\nabla f(x^i)\|_* + \tfrac{L}{2}\alpha_i^2 \ , \end{aligned} \tag{31}$$

where the equality follows since $\|\nabla f(x^k)\|_* = \max_{d:\|d\| \leq 1} \nabla f(x^k)^T d = \nabla f(x^k)^T d^k$. Summing the above for $i = 0, \ldots, k$ yields:

$$f^* \leq f(x^{k+1}) \leq f(x^0) - \textstyle\sum_{i=0}^k \alpha_i \|\nabla f(x^i)\|_* + \frac{L}{2}\sum_{i=0}^k \alpha_i^2 \ . \tag{32}$$

Next notice that

$$\sum_{i=0}^k \alpha_i \|\nabla f(x^i)\|_* \geq \left(\sum_{i=0}^k \alpha_i\right) \left(\min_{i \in \{0,\ldots,k\}} \|\nabla f(x^i)\|_*\right) \ ,$$

and substituting this inequality above and rearranging yields

$$\min_{i \in \{0,\ldots,k\}} \|\nabla f(x^i)\|_* \leq \frac{f(x^0) - f^* + \frac{L}{2}\sum_{i=0}^k \alpha_i^2}{\sum_{i=0}^k \alpha_i} \ . \tag{33}$$

Now suppose we use the step-sizes (17). Substituting (17) into (31) yields:

$$f(x^{i+1}) \le f(x^i) - \tfrac{1}{2L}\|\nabla f(x^i)\|_*^2 \ , \tag{34}$$

which shows that the values $f(x^i)$ are monotone decreasing and hence $f(x^i) \le f(x^0)$, whereby $x^i \in \mathcal{S}_0$. Substituting the step-sizes (17) into (32) yields after rearranging:

$$\sum_{i=0}^{k} \|\nabla f(x^i)\|_*^2 \ \le 2L(f(x^0) - f(x^{k+1})) \le \ 2L(f(x^0) - f^*) \ , \tag{35}$$

and therefore

$$(k+1)\left(\min_{i\in\{0,\ldots,k\}} \|\nabla f(x^i)\|_*\right)^2 \le \sum_{i=0}^{k} \|\nabla f(x^i)\|_*^2 \ \le \ 2L(f(x^0) - f^*) \ ,$$

and rearranging yields *(iv)*. Now suppose as well that Dist_0 is finite, and let x^i be an iterate of the steepest descent method. It was shown above that $x^i \in \mathcal{S}_0$, whereby there exists $x^* \in \mathcal{S}^*$ for which $\|x^i - x^*\| \le \mathrm{Dist}_0$, and from the gradient inequality for the convex function $f(\cdot)$ it holds that

$$\begin{aligned} f^* = f(x^*) \ &\ge \ f(x^i) + \nabla f(x^i)^T(x^* - x^i) \\ &\ge \ f(x^i) - \|\nabla f(x^i)\|_* \|x^* - x^i\| \\ &\ge \ f(x^i) - \|\nabla f(x^i)\|_* \mathrm{Dist}_0 \ , \end{aligned}$$

and rearranging the above yields $\|\nabla f(x^i)\|_* \ge \frac{f(x^i)-f^*}{\mathrm{Dist}_0}$. Substituting this inequality into (34) and subtracting f^* from both sides yields:

$$f(x^{i+1}) - f^* \le f(x^i) - f^* - \frac{(f(x^i) - f^*)^2}{2L\mathrm{Dist}_0^2} \ .$$

Define $a_i := f(x^i) - f^*$, and it follows that the nonnegative series $\{a_i\}$ satisfies $a_{i+1} \le a_i - \frac{a_i^2}{2L\mathrm{Dist}_0^2}$. A standard induction argument (see for example Lemma 3.5 of [5]) then establishes that

$$a_k \le \frac{1}{\frac{1}{a_0} + \frac{k}{2L\mathrm{Dist}_0^2}} \ ,$$

which when rearranged yields the first inequality of *(i)*. The second inequality of *(i)* follows since $\hat{K}^0 > 0$.

Rearranging (34) yields $\|\nabla f(x^i)\|_*^2 \le 2L(f(x^i) - f(x^{i+1}) \le 2L(f(x^i) - f^*)$, which after taking square roots proves the first inequality of *(ii)*, and the second inequality of *(ii)* follows by substituting in the bound on $f(x^k) - f^*$ from the first inequality of *(i)*.

To prove *(iii)*, use $k-1$ in (35), and use the step-lengths (17) to yield:

$$2L(f(x^0) - f^*) \ge \sum_{i=0}^{k-1} \|\nabla f(x^i)\|_*^2 = L^2 \sum_{i=0}^{k-1} \alpha_i^2 \ge \left(\frac{1}{k}\right) L^2 \left(\sum_{i=0}^{k-1} \alpha_i\right)^2 \ge \left(\frac{1}{k}\right) L^2 \|x^k - x^0\|^2 \ ,$$

and rearranging the above yields *(iii)*. □

A.3 Proof of Proposition 3.1

We first present a property of the following "prox" function $d(\cdot) : [0,1]^n \to \mathbb{R}$ defined by:

$$d(w) := \frac{1}{n}\left[\sum_{i=1}^{n} w_i \ln(w_i) + (1-w_i)\ln(1-w_i)\right] \ , \tag{36}$$

where $\alpha \ln(\alpha) := 0$ for $\alpha = 0$.

Proposition A.1. *Consider the function $d(\cdot) : [0,1]^n \to \mathbb{R}$ given by (36). It holds that $d(\cdot)$ is a $\sigma := \frac{4}{n}$-strongly convex function with respect to the Euclidean norm $\|w\| := \|w\|_2$.*

Proof: Let $G := [0,1]^n$, consider any point $w \in \text{int}G$, and let $H(w)$ denote the Hessian matrix of $d(\cdot)$ at w. The off-diagonal components of $H(w)$ are all zero, and the i^{th} diagonal component is $H_{ii}(w) = \frac{1}{n}\frac{1}{w_i(1-w_i)} \ge \frac{4}{n}$, and hence $v^T[H(w)]v \ge \frac{4}{n}v^Tv$ for any v. Now let $y \in \text{int}G$. Invoking an intermediate value theorem of calculus, there exists a scalar $c \in [0,1]$ for which it holds that:

$$d(y) = d(w) + \nabla d(w)^T(y-w) + \tfrac{1}{2}(y-w)H(w + c(y-w))(y-w) \ ,$$

whereby:

$$d(y) \ge d(w) + \nabla d(w)^T(y-w) + \tfrac{1}{2}(y-w)\left[\tfrac{4}{n}I\right](y-w) = d(w) + \nabla d(w)^T(y-w) + \tfrac{1}{2}\tfrac{4}{n}\|y-w\|_2^2 \ .$$

This proves that $d(\cdot)$ is $\sigma = \frac{4}{n}$-strongly convex on $\text{int}G$, and a continuity argument establishes the result for all of G. □

Proof of Proposition 3.1: We first claim that

$$L_n(\beta) = \max_{w \in [0,1]^n} \left\{-w^T[\tfrac{1}{n}Y\mathbf{X}]\beta - d(w)\right\} \tag{37}$$

where $d(\cdot)$ is given by (36), and the unique optimal solution to the maximization problem in (37) is:

$$w^*(\beta)_i = \frac{1}{1+\exp(y_i\beta^T\mathbf{x}_i)} \ \ , \ i = 1,\ldots,n \ . \tag{38}$$

Indeed, it is easy to verify through optimality conditions that the unique optimal solution to the maximization problem in (37) is given by (38), and direct substitution and simplification of terms then yields the equality in (37). Using the representation (37), Theorem 1 of [23] implies that the Lipschitz constant L of the gradient of $L_n(\cdot)$ is at most $[\frac{1}{n}\|\mathbf{X}\|_{\cdot,2}]^2/\sigma$ where σ is the strong convexity parameter of the function $d(\cdot)$. From Proposition A.1 it holds that $\sigma \ge \frac{4}{n}$, which implies that $L \le [\frac{1}{n}\|\mathbf{X}\|_{\cdot,2}]^2/\sigma \le [\frac{1}{4n}]\|\mathbf{X}\|_{\cdot,2}^2$. □

A.4 Proof of Theorem 3.3

Following Bach [2, 3], a three-times differentiable convex function $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ is said to be generalized self-concordant (with respect to the norm $\|\cdot\|$) if there is a constant $R > 0$ such that for all $x, \hat{x} \in \mathbb{R}^p$, the scalar function $\varphi(\cdot) : t \mapsto f(x + t(\hat{x} - x))$ satisfies:

$$|\varphi'''(t)| \ \le \ R\|x - \hat{x}\|\varphi''(t) \ \text{ for all } t \in \mathbb{R} \ . \tag{39}$$

The above definition is a slight modification of that given in [2, 3], which works with the ℓ_2 norm. The following proposition, which is a very minor generalization of a result shown in [2], demonstrates that the logistic loss function $L_n(\cdot)$ is generalized self-concordant with constant $R = \|\mathbf{X}\|_{\cdot,\infty}$.

Proposition A.2. *The logistic loss function $L_n(\cdot)$ is generalized self-concordant (with respect to the norm $\|\cdot\|$) with constant $R = \|\mathbf{X}\|_{\cdot,\infty}$.*

Proof: Let $\beta, \hat{\beta} \in \mathbb{R}^p$ be given and define the scalar function $\varphi(\cdot) : \mathbb{R} \to \mathbb{R}$ by $\varphi(t) := L_n(\beta + t(\hat{\beta} - \beta))$. A simple calculation yields:

$$\begin{aligned}
|\varphi'''(t)| &= \left| \frac{1}{n} \sum_{i=1}^{n} \ell'''(y_i\beta^T\mathbf{x}_i + ty_i(\hat{\beta} - \beta)^T\mathbf{x}_i) \cdot (y_i(\hat{\beta} - \beta)^T\mathbf{x}_i)^3 \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} |\ell'''(y_i\beta^T\mathbf{x}_i + ty_i(\hat{\beta} - \beta)^T\mathbf{x}_i)| \cdot (y_i(\hat{\beta} - \beta)^T\mathbf{x}_i)^2 \cdot |y_i(\hat{\beta} - \beta)^T\mathbf{x}_i| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \ell''(y_i\beta^T\mathbf{x}_i + ty_i(\hat{\beta} - \beta)^T\mathbf{x}_i) \cdot (y_i(\hat{\beta} - \beta)^T\mathbf{x}_i)^2 \cdot |y_i(\hat{\beta} - \beta)^T\mathbf{x}_i| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \ell''(y_i\beta^T\mathbf{x}_i + ty_i(\hat{\beta} - \beta)^T\mathbf{x}_i) \cdot (y_i(\hat{\beta} - \beta)^T\mathbf{x}_i)^2 \cdot \|\mathbf{x}_i\|_*\|\beta - \hat{\beta}\| \\
&\leq \frac{\|\mathbf{X}\|_{\cdot,\infty}\|\beta - \hat{\beta}\|}{n} \sum_{i=1}^{n} \ell''(y_i\beta^T\mathbf{x}_i + ty_i(\hat{\beta} - \beta)^T\mathbf{x}_i) \cdot (y_i(\hat{\beta} - \beta)^T\mathbf{x}_i)^2 \\
&= \|\mathbf{X}\|_{\cdot,\infty}\|\beta - \hat{\beta}\|\varphi''(t) \ ,
\end{aligned}$$

where the second inequality above uses $|\ell'''(\cdot)| \leq \ell''(\cdot)$, the third uses Hölder's inequality, and the final inequality uses $\|\mathbf{X}\|_{\cdot,\infty} = \max_{i\in\{1,\ldots,n\}} \|\mathbf{x}_i\|_*$. □

In order to prove Theorem 3.3, we use the following lemma which is a minor extension of Lemma 9 of [3].

Lemma A.1. (essentially Bach [3], Lemma 9) *Suppose that* DegNSEP$^* > 0$. *Let β satisfying $L_n(\beta) \leq \ln(2)$ be given, and let β^* be the unique optimal solution of LR. Then it holds that:*

(i) $\quad L_n(\beta) - L_n^* \ \leq \ \left(1 + \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right) \frac{\|\nabla L_n(\beta)\|_*^2}{\nu^*(H(\beta^*))}$, *and*

(ii) $\quad \|\beta - \beta^*\| \ \leq \ \left(1 + \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right) \left(\frac{\|\mathbf{X}\|_{\cdot,2}}{\nu^*(H(\beta^*))}\right) \sqrt{\frac{L_n(\beta) - L_n^*}{2n}}$.

If in addition β satisfies $\frac{\|\nabla L_n(\beta)\|_*\|\mathbf{X}\|_{\cdot,\infty}}{\nu^*(H(\beta^*))} \ \leq \ \frac{3}{4}$, then it holds that:

(iii) $\quad L_n(\beta) - L_n^* \ \leq \ \frac{2\|\nabla L_n(\beta)\|_*^2}{\nu^*(H(\beta^*))}$, *and*

(iv) $\quad \|\beta - \beta^*\| \ \leq \ \frac{\|\mathbf{X}\|_{\cdot,2}}{\nu^*(H(\beta^*))} \sqrt{\frac{2(L_n(\beta) - L_n^*)}{n}}$.

Proof: First note that if $\beta = \beta^*$ then the lemma is trivial, so we assume that $\beta \neq \beta^*$. We will apply Lemma 13 of [3] to the scalar function $\varphi(\cdot) : [0, 1] \to \mathbb{R}$ defined by $\varphi(t) := L_n(\beta^* + t(\beta - \beta^*))$. Let us define $S := \|\mathbf{X}\|_{\cdot,\infty}\|\beta - \beta^*\|$, and note that Proposition A.2 implies that $\varphi(\cdot)$ satisfies

$|\varphi'''(t)| \le S\varphi''(t)$ for all $t \in [0,1]$. Simple calculations yield:

$\varphi'(t) = \nabla L_n(\beta^*+t(\beta-\beta^*))^T(\beta-\beta^*)$ and $\varphi''(t) = (\beta-\beta^*)^T H(\beta^*+t(\beta-\beta^*))(\beta-\beta^*)$ for all $t \in [0,1]$.

In particular, we have $\varphi'(0) = 0$ by the optimality of β^* and

$$\varphi'(1) = \nabla L_n(\beta)^T(\beta - \beta^*) \le \|\nabla L_n(\beta)\|_* \|\beta - \beta^*\| \ , \tag{40}$$

by Hölder's inequality. Moreover, we have:

$$\varphi''(0) = (\beta - \beta^*)^T H(\beta^*)(\beta - \beta^*) \ge \nu^*(H(\beta^*))\|\beta - \beta^*\|^2 > 0 \ , \tag{41}$$

by the definition of $\nu^*(H(\beta^*))$ in (7), part *(ii)* of Proposition 2.1, and since $\beta \ne \beta^*$. Therefore $\varphi(\cdot)$ satisfies the hypotheses of Lemma 13 of [3], and a direct application of this lemma yields:

$$\frac{\varphi'(1)}{\varphi''(0)} S \ge 1 - \exp(-S) \ , \ \text{and} \ \ \varphi(1) \le \varphi(0) + \frac{\varphi'(1)^2}{\varphi''(0)}(1+S) \ . \tag{42}$$

Following the proof of Proposition 2.1, we have that $S = \|\mathbf{X}\|_{\cdot,\infty}\|\beta - \beta^*\| \le \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}$ since β satisfies $L_n(\beta) \le \ln(2)$. Substituting this upper bound on S along with the inequalities (40) and (41) into the rightmost inequality in (42) yields part *(i)* of the lemma. Making the same substitutions into the leftmost inequality in (42) and rearranging yields:

$$\|\beta - \beta^*\| \le \frac{S\|\nabla L_n(\beta)\|_*}{(1-\exp(-S))\nu^*(H(\beta^*))} \le \left(1 + \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right)\frac{\|\nabla L_n(\beta)\|_*}{\nu^*(H(\beta^*))} \ ,$$

where the second inequality uses $\frac{S}{1-\exp(-S)} \le 1 + S$. Applying (34) along with Proposition 3.1 in this context yields $L_n^* \le L_n(\beta) - \frac{2n}{\|\mathbf{X}\|_{\cdot,2}^2}\|\nabla L_n(\beta)\|_*^2$, which after rearranging terms and combining with the above inequality yields part *(ii)* of the lemma.

Now suppose that $\frac{\|\nabla L_n(\beta)\|_*\|\mathbf{X}\|_{\cdot,\infty}}{\nu^*(H(\beta^*))} \le \frac{3}{4}$ additionally holds. Then:

$$\frac{\varphi'(1)S}{\varphi''(0)} \le \frac{\|\nabla L_n(\beta)\|_*\|\beta - \beta^*\|^2\|\mathbf{X}\|_{\cdot,\infty}}{\nu^*(H(\beta^*))\|\beta - \beta^*\|^2} = \frac{\|\nabla L_n(\beta)\|_*\|\mathbf{X}\|_{\cdot,\infty}}{\nu^*(H(\beta^*))} \le \frac{3}{4} \ .$$

Hence following Lemma 13 of [3], it holds that $\varphi''(0) \le 2\varphi'(1)$ and $\varphi(1) \le \varphi(0) + 2\frac{\varphi'(1)^2}{\varphi''(0)}$, and making the same substitutions as before yields parts *(iii)* and *(iv)*. □

Proof of Theorem 3.3: Let $k \ge 0$ be given. Applying (34) along with Proposition 3.1 in this context yields:

$$L_n(\beta^{k+1}) \le L_n(\beta^k) - \tfrac{2n}{\|\mathbf{X}\|_{\cdot,2}^2}\|\nabla L_n(\beta^k)\|_*^2 \ . \tag{43}$$

In particular, $L_n(\beta^{k+1}) \le L_n(\beta^k)$, which implies that $L_n(\beta^i) \le L_n(\beta^0) = \ln(2)$ for all $i \ge 0$. Therefore, we may apply item *(i)* of Lemma A.1, which yields:

$$L_n(\beta^k) - L_n^* \ \le \ \left(1 + \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right)\frac{\|\nabla L_n(\beta^k)\|_*^2}{\nu^*(H(\beta^*))} \ .$$

Combining the above inequality with (43) yields:

$$L_n(\beta^{k+1}) \le L_n(\beta^k) - \frac{2n\nu^*(H(\beta^*))(L_n(\beta^k) - L_n^*)}{\|\mathbf{X}\|_{\cdot,2}^2 \left(1 + \frac{2\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\text{DegNSEP}^*}\right)} .$$

Finally, subtracting L_n^* from both sides of the above and rearranging terms yields:

$$L_n(\beta^{k+1}) - L_n^* \le (L_n(\beta^k) - L_n^*)\left(1 - \frac{2(\text{DegNSEP}^*)\nu^*(H(\beta^*))n}{(\text{DegNSEP}^* + 2\ln(2)\|\mathbf{X}\|_{\cdot,\infty})\|\mathbf{X}\|_{\cdot,2}^2}\right) ,$$

which immediately implies part *(i)* of the theorem. Part *(ii)* of the theorem follows by substituting the bound on $L_n(\beta^k) - L_n^*$ from part *(i)* of the theorem into the bound on $\|\beta^k - \beta^*\|$ from part *(ii)* of Lemma A.1.

Now assume that $k \ge \check{K}$. Part *(iii)* of Theorem 3.2 implies that:

$$\frac{\|\nabla L_n(\beta^k)\|_*\|\mathbf{X}\|_{\cdot,\infty}}{\nu^*(H(\beta^*))} \;\le\; \frac{\|\mathbf{X}\|_{\cdot,2}^2 \ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\nu^*(H(\beta^*))\sqrt{k} \cdot n \cdot \text{DegNSEP}^*} \le \frac{3}{4} ,$$

where the final inequality uses the fact that $k \ge \check{K} \ge \frac{16\ln(2)^2\|\mathbf{X}\|_{\cdot,2}^4\|\mathbf{X}\|_{\cdot,\infty}^2}{9n^2(\text{DegNSEP}^*)^2\nu^*(H(\beta^*))^2}$. Thus we may apply part *(iii)* of Lemma A.1, which yields $L_n(\beta^k) - L_n^* \le \frac{2\|\nabla L_n(\beta^k)\|_*^2}{\nu^*(H(\beta^*))}$. By the same arguments as above, we obtain:

$$L_n(\beta^{k+1}) - L_n^* \le (L_n(\beta^k) - L_n^*)\left(1 - \frac{\nu^*(H(\beta^*))n}{\|\mathbf{X}\|_{\cdot,2}^2}\right) ,$$

which immediately implies part *(iii)* of the thoerem. Part *(iv)* of the theorem similarly follows by substituting the bound on $L_n(\beta^k) - L_n^*$ from part *(iii)* into the bound on $\|\beta^k - \beta^*\|$ from part *(iv)* of Lemma A.1. □

A.5 Proofs of Lemmas 3.1 and 4.1

Denote the univariate logistic loss function by $\ell(t) := \ln(1 + \exp(-t))$. We start with the following quite general proposition which presents a bound on the iterate sequence $\{\beta_k\}$ of *any* algorithm whose step direction g_k is an average of gradients of the logistic loss function over a subset S_k of the observations, for all k.

Proposition A.3. *Consider any algorithm for solving the logistic regression problem* (2), *and let* $\{\beta^k\}$ *denote the iterate sequence. Suppose that* $\beta^0 := 0$ *and* $\{\beta^k\}$ *satisfies:*

$$\beta^{k+1} = \beta^k - \alpha_k g_k \quad \textit{where} \quad g_k = \frac{1}{|S_k|}\sum_{i \in S_k} \nabla_\beta \ell(y_i(\beta^k)^T\mathbf{x}_i) \tag{44}$$

for some $S_k \subseteq \{1, \ldots, n\}$*, for all* $k \ge 0$*. If* $\alpha_j \le \frac{2}{\|\mathbf{X}\|_{2,\infty}^2}$ *for all* $j \ge 0$*, then for all* $\beta \in \mathbb{R}^p$ *and for all* $k \ge 0$ *it holds that:*

$$\|\beta^k - \beta\|_2^2 \;\le\; \|\beta\|_2^2 + 2\sum_{j=0}^{k-1}\frac{\alpha_j}{|S_j|}\sum_{i \in S_j}\ell(y_i\beta^T\mathbf{x}_i) .$$

Proof: For any $j \in \{0, \dots, k-1\}$ it holds that:

$$\begin{aligned}
\|\beta^{j+1} - \beta\|_2^2 &= \|\beta^j - \beta - \alpha_j g_j\|_2^2 \\
&= \|\beta^j - \beta\|_2^2 - 2\alpha_j g_j^T(\beta^j - \beta) + \alpha_j^2\|g_j\|_2^2 \\
&\leq \|\beta^j - \beta\|_2^2 + 2\tfrac{\alpha_j}{|S_j|}\textstyle\sum_{i\in S_j}[\ell(y_i\beta^T\mathbf{x}_i) - \ell(y_i(\beta^j)^T\mathbf{x}_i)] + \alpha_j^2\|g_j\|_2^2 \\
&= \|\beta^j - \beta\|_2^2 + 2\tfrac{\alpha_j}{|S_j|}\textstyle\sum_{i\in S_j}[\ell(y_i\beta^T\mathbf{x}_i) - \ell(y_i(\beta^j)^T\mathbf{x}_i)] + \tfrac{\alpha_j^2}{|S_j|^2}\|\textstyle\sum_{i\in S_j}\nabla_\beta\ell(y_i(\beta^j)^T\mathbf{x}_i)\|_2^2 \\
&\leq \|\beta^j - \beta\|_2^2 + 2\tfrac{\alpha_j}{|S_j|}\textstyle\sum_{i\in S_j}[\ell(y_i\beta^T\mathbf{x}_i) - \ell(y_i(\beta^j)^T\mathbf{x}_i)] + \tfrac{\alpha_j^2}{|S_j|^2}\left(\textstyle\sum_{i\in S_j}\|\nabla_\beta\ell(y_i(\beta^j)^T\mathbf{x}_i)\|_2\right)^2 \\
&\leq \|\beta^j - \beta\|_2^2 + 2\tfrac{\alpha_j}{|S_j|}\textstyle\sum_{i\in S_j}[\ell(y_i\beta^T\mathbf{x}_i) - \ell(y_i(\beta^j)^T\mathbf{x}_i)] + \tfrac{\alpha_j^2}{|S_j|}\textstyle\sum_{i\in S_j}\|\nabla_\beta\ell(y_i(\beta^j)^T\mathbf{x}_i)\|_2^2 \\
&= \|\beta^j - \beta\|_2^2 + 2\tfrac{\alpha_j}{|S_j|}\textstyle\sum_{i\in S_j}\ell(y_i\beta^T\mathbf{x}_i) + \tfrac{\alpha_j}{|S_j|}\textstyle\sum_{i\in S_j}[\alpha_j\|\nabla_\beta\ell(y_i(\beta^j)^T\mathbf{x}_i)\|_2^2 - 2\ell(y_i(\beta^j)^T\mathbf{x}_i)] \ ,
\end{aligned}$$

where the first inequality above is an application of the gradient inequality, the second inequality above uses the triangle inequality, and the third inequality utilizes an inequality between the ℓ_1 and ℓ_2 norms. Now since $\alpha_j \leq \frac{2}{\|\mathbf{X}\|_{2,\infty}^2}$, it holds that:

$$\alpha_j\|\nabla_\beta\ell(y_i(\beta^j)^T\mathbf{x}_i)\|_2^2 = \alpha_j\ell'(y_i(\beta^j)^T\mathbf{x}_i)^2\|\mathbf{x}_i\|_2^2 \leq 2\ell'(y_i(\beta^j)^T\mathbf{x}_i)^2 \leq 2\ell(y_i(\beta^j)^T\mathbf{x}_i)) \ ,$$

where the last inequality uses $\ell'(\cdot)^2 \leq |\ell'(\cdot)| \leq \ell(\cdot)$. Therefore:

$$\|\beta^{j+1} - \beta\|_2^2 \ \leq \ \|\beta^j - \beta\|_2^2 + 2\frac{\alpha_j}{|S_j|}\sum_{i\in S_j}\ell(y_i\beta^T\mathbf{x}_i) \ ,$$

and summing the previous inequality over $j \in \{0, \dots, k-1\}$ yields the result. □

Proofs of Lemmas 3.1 and 4.1: We present a unified proof of these two results. Let $\bar{\beta}$ denote the normalized maximum margin hyperplane, i.e., the optimal solution of (8), and define $\tilde{\beta}^k := (\ln(k)/\mathrm{DegSEP}^*)\bar{\beta}$. For each $i \in \{1, \dots, n\}$, we have:

$$y_i(\tilde{\beta}^k)^T\mathbf{x}_i = (\ln(k)/\mathrm{DegSEP}^*)y_i\bar{\beta}^T\mathbf{x}_i \geq (\ln(k)/\mathrm{DegSEP}^*)\rho(\bar{\beta}) = \ln(k) \ .$$

Furthermore, since $\ell(t) \leq \exp(-t)$, it holds that:

$$\ell(y_i(\tilde{\beta}^k)^T\mathbf{x}_i) \leq \ell(\ln(k)) \leq \exp(-\ln(k)) = 1/k \ . \tag{45}$$

Clearly, the conditions for Proposition A.3 are satisfied by ℓ_2 steepest descent under the assumptions of Lemma 3.1 (wherein $|S_k| = n$ for all k) as well as SGD under the assumptions of Lemma 4.1 (wherein $|S_k| = 1$ for all k). (Note that in this proof α_j refers to the step-size with respect to the unnormalized version of ℓ_2 steepest descent.) Therefore, in both cases we may apply Proposition A.3 using $\beta = \tilde{\beta}^k$ to yield:

$$\|\beta^k - \tilde{\beta}^k\|_2^2 \ \leq \ \|\tilde{\beta}^k\|_2^2 + 2\sum_{j=0}^{k-1}\frac{\alpha_j}{|S_j|}\sum_{i\in S_j}\ell(y_i(\tilde{\beta}^k)^T\mathbf{x}_i) \ \leq \ \frac{\ln(k)^2}{(\mathrm{DegSEP}^*)^2} + \frac{4}{\|\mathbf{X}\|_{2,\infty}^2} \ ,$$

where the final inequality uses (45) as well as $\alpha_j \leq \frac{2}{\|\mathbf{X}\|_{2,\infty}^2}$. Therefore:

$$\|\beta^k\| \leq \|\tilde{\beta}^k\|_2 + \|\beta^k - \tilde{\beta}^k\|_2 \leq \frac{\ln(k)}{\mathrm{DegSEP}^*} + \sqrt{\frac{\ln(k)^2}{(\mathrm{DegSEP}^*)^2} + \frac{4}{\|\mathbf{X}\|_{2,\infty}^2}} \leq \frac{2\ln(k)}{\mathrm{DegSEP}^*} + \frac{2}{\|\mathbf{X}\|_{2,\infty}} \ .$$

□

A.6 Proof of Theorem 4.1

To prove *(i)*, let $x \in \mathbb{R}^p$ be fixed and notice that for each $i \geq 0$ it holds that:

$$\|x^{i+1} - x\|_2^2 = \|x^i - \bar{\alpha}\tilde{\nabla} f(x^i) - x\|_2^2 = \|x^i - x\|_2^2 - 2\bar{\alpha}\tilde{\nabla} f(x^i)^T(x^i - x) + \bar{\alpha}^2\|\tilde{\nabla} f(x^i)\|_2^2$$

Rearranging terms, summing over $i \in \{0, \ldots, k\}$, and dividing by $2\bar{\alpha}(k+1)$ yields:

$$\begin{aligned} \frac{1}{k+1}\sum_{i=0}^{k} \tilde{\nabla} f(x^i)^T(x^i - x) &= \frac{\|x^0 - x\|_2^2}{2\bar{\alpha}(k+1)} - \frac{\|x^{k+1} - x\|_2^2}{2\bar{\alpha}(k+1)} + \frac{\bar{\alpha}}{2(k+1)}\sum_{i=0}^{k}\|\tilde{\nabla} f(x^i)\|_2^2 \\ &\leq \frac{\|x^0 - x\|_2^2}{2\bar{\alpha}(k+1)} + \frac{\bar{\alpha}}{2(k+1)}\sum_{i=0}^{k}\|\tilde{\nabla} f(x^i)\|_2^2 \ . \end{aligned} \tag{46}$$

Now, by the law of iterated expectations, for each $i \in \{0, \ldots, k\}$ it holds that

$$\begin{aligned} \mathbb{E}\left[\tilde{\nabla} f(x^i)^T(x^i - x)\right] &= \mathbb{E}\left[\mathbb{E}[\tilde{\nabla} f(x^i)^T(x^i - x) \mid x_i]\right] \\ &= \mathbb{E}\left[\mathbb{E}[\tilde{\nabla} f(x^i) \mid x_i]^T(x^i - x)\right] \\ &= \mathbb{E}\left[\nabla f(x^i)^T(x^i - x)\right] \ , \end{aligned}$$

where the last equality follows from the definition of the stochastic gradient, i.e., $\mathbb{E}[\tilde{\nabla} f(x^i) \mid x_i] = \nabla f(x^i)$. After taking the expectation of both sides of (46) and combining with the above we obtain

$$\begin{aligned} \frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[\nabla f(x^i)^T(x^i - x)\right] &\leq \frac{\|x^0 - x\|_2^2}{2\bar{\alpha}(k+1)} + \frac{\bar{\alpha}}{2(k+1)}\sum_{i=0}^{k}\mathbb{E}[\|\tilde{\nabla} f(x^i)\|_2^2] \\ &\leq \frac{\|x^0 - x\|_2^2}{2\bar{\alpha}(k+1)} + \frac{\bar{\alpha}M^2}{2} \ , \end{aligned}$$

where the second inequality follows from (22). The gradient equality (which holds for each realization of x^i) states that $f(x^i) - f(x) \leq \nabla f(x^i)^T(x^i - x)$, and averaging the expectation of these inequalities over $i \in \{0, \ldots, k\}$ yields

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[f(x^i)\right] - f(x) \leq \frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[\nabla f(x^i)^T(x^i - x)\right] \ . \tag{47}$$

Finally, in the case of Option A, Jensen's inequality implies *(i)* and, in the case of Option B, another iterated expectations argument implies that $\mathbb{E}[f(\hat{x}^k)] = \frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[f(x^i)\right]$ from which *(i)* directly follows.

Item *(ii)* follows directly from the format of the updates in Step (2.) of Algorithm 2 as well as the triangle inequality. To prove *(iii)*, we use the smoothness of the objective function. In particular, applying (30) to the iterates of Algorithm 2 yields for each $i \in \{0, \ldots, k\}$:

$$\begin{aligned} f(x^{i+1}) &\leq f(x^i) + \nabla f(x^i)^T(x^{i+1} - x^i) + \tfrac{L}{2}\|x^{i+1} - x^i\|_2^2 \\ &= f(x^i) - \bar{\alpha}\nabla f(x^i)^T\tilde{\nabla} f(x^i) + \tfrac{L\bar{\alpha}^2}{2}\|\tilde{\nabla} f(x^i)\|_2^2 \ . \end{aligned} \tag{48}$$

Notice that the law of iterated expectations as well as the definition of the stochastic gradient yields:

$$\begin{aligned}\mathbb{E}\left[\nabla f(x^i)^T\tilde{\nabla} f(x^i)\right] &= \mathbb{E}\left[\mathbb{E}[\nabla f(x^i)^T\tilde{\nabla} f(x^i) \mid x^i]\right] \\ &= \mathbb{E}\left[\nabla f(x^i)^T\mathbb{E}[\tilde{\nabla} f(x^i) \mid x_i]\right] \\ &= \mathbb{E}\left[\nabla f(x^i)^T\nabla f(x^i)\right] \\ &= \mathbb{E}\left[\|\nabla f(x^i)\|_2^2\right] \ .\end{aligned}$$

Therefore, taking the expectation of both sides of (48) and using (22) yields:

$$\mathbb{E}[f(x^{i+1})] \ \le \ \mathbb{E}[f(x^i)] - \bar{\alpha}\cdot\mathbb{E}\left[\|\nabla f(x^i)\|_2^2\right] + \frac{L\bar{\alpha}^2M^2}{2} \ .$$

Rearranging terms and summing over $i \in \{0,\ldots,k\}$ yields:

$$\bar{\alpha}\sum_{i=0}^{k}\mathbb{E}\left[\|\nabla f(x^i)\|_2^2\right] \ \le \ f(x^0) - \mathbb{E}[f(x^{k+1})] + \frac{L\bar{\alpha}^2M^2(k+1)}{2} \ .$$

Then using $f^* \le \mathbb{E}[f(x^{k+1})]$ and dividing by $\bar{\alpha}(k+1)$ yields:

$$\frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[\|\nabla f(x^i)\|_2^2\right] \ \le \ \frac{f(x^0)-f^*}{\bar{\alpha}(k+1)} + \frac{\bar{\alpha}LM^2}{2} \ .$$

Finally, since we are in the case of Option B, another iterated expectations argument implies that $\mathbb{E}\left[\|\nabla f(\hat{x}^k)\|_2^2\right] = \frac{1}{k+1}\sum_{i=0}^{k}\mathbb{E}\left[\|\nabla f(x^i)\|_2^2\right]$ from which *(iii)* directly follows. □

A.7 Proof of Proposition 4.2

Recall that the scalar logistic loss function is denoted by $\ell(\cdot) : \mathbb{R} \to \mathbb{R}$, which is defined by $\ell(t) := \ln(1+\exp(-t))$. A simple calculation shows that $\ell''(t) = \frac{\exp(t)}{(\exp(t)+1)^2} \le \frac{1}{4}$ for all $t \in \mathbb{R}$. As mentioned in Section 4.2, it follows from item (2.) of Assumption 4.1 that $L_{\mathcal{D}}(\cdot)$ is continuous, convex, differentiable, and satisfies $\nabla L_{\mathcal{D}}(\beta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\nabla_\beta \ell(y\beta^T\mathbf{x})\right] = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell'(y\beta^T\mathbf{x})\cdot y\mathbf{x}\right]$ for all $\beta \in \mathbb{R}^p$ (see, e.g., Section 7.2.4 of [29]). Moreover, it can also be demonstrated (again by Section 7.2.4 of [29]) that $L_{\mathcal{D}}(\cdot)$ is twice differentiable. Letting $H(\beta)$ denote the Hessian matrix of $L_{\mathcal{D}}(\beta)$ at β, then it holds that $H(\beta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell''(y\beta^T\mathbf{x})\mathbf{x}\mathbf{x}^T\right]$.

Recall that a twice differentiable convex function is L-smooth with respect to the ℓ_2 norm on $\mathbb{R}^p$ if and only if $H(\beta) \preceq LI_p$ for all $\beta \in \mathbb{R}^p$, where I_p denotes the $p\times p$ identity matrix. Now, for any $(\mathbf{x}, y)$, since $\ell''(y\beta^T\mathbf{x}) \le \frac{1}{4}$, we have that $\ell''(y\beta^T\mathbf{x})\mathbf{x}\mathbf{x}^T \preceq \frac{1}{4}\mathbf{x}\mathbf{x}^T$. Therefore, it holds that

$$H(\beta) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell''(y\beta^T\mathbf{x})\mathbf{x}\mathbf{x}^T\right] \preceq \tfrac{1}{4}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\mathbf{x}\mathbf{x}^T\right] = \tfrac{1}{4}\Sigma \preceq \tfrac{1}{4}\lambda_{\max}(\Sigma)I_p \ ,$$

which demonstrates that $L_{\mathcal{D}}(\cdot)$ is $\frac{1}{4}\lambda_{\max}(\Sigma)$-smooth.

□

A.8 Proof of Proposition 4.3

Again, recalling the notation $\ell(t) := \ln(1 + \exp(-t))$, note that $\ell'(t) = -\frac{1}{\exp(t)+1} \in (-1, 0)$ for all $t \in \mathbb{R}$. Then the stochastic gradient is $\nabla_\beta \ell(y\beta^T\mathbf{x})$ where $(\mathbf{x}, y) \sim \mathcal{D}$ and it holds that

$$\begin{aligned}
\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|\nabla_\beta \ell(y\beta^T\mathbf{x})\|_2^2\right] &= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|\ell'(y\beta^T\mathbf{x}) \cdot y\mathbf{x}\|_2^2\right] \\
&= \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[|\ell'(y\beta^T\mathbf{x})| \cdot \|\mathbf{x}\|_2^2\right] \\
&\leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\|\mathbf{x}\|_2^2\right] \\
&= \mathrm{Tr}(\Sigma) \ .
\end{aligned}$$

□