May 2020

Worst Case Datasets for Solving Binary Logistic Regression via Deterministic First-Order Methods

Trevor Squires
Clemson University, tsquire@clemson.edu

Worst Case Datasets for Solving Binary Logistic Regression via Deterministic First-Order Methods

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Mathematical Sciences

by
Trevor Squires
May 2020

Accepted by:
Dr. Yuyuan Ouyang, Committee Chair
Dr. Fei Xue
Dr. Christopher McMahan

In this thesis, we construct worst case binary logistic regression datasets for any deterministic first order methods. We show that our datasets require at least $\mathcal{O}(1/\sqrt{\varepsilon})$ first-order oracle inquires to obtain a $\varepsilon-$approximate solution under the assumption that the problem dimension is sufficiently large. Using our result, on worst case datasets we conclude that existing algorithms such as Nesterov's Optimal Gradient Descent are optimal algorithms for solving binary logistic regression under large scale assumptions. Our analysis combines Nemirovski's Krylov subspace technique and Nesterov's construction of worst case convex quadratic programming problem instance. Our result is the first worst case dataset constructed against all first order methods for solving binary logistic regression, and a new worst case instance among smooth convex optimization problems.

Acknowledgments

This research is supported by National Science Foundation grant DMS-1913006.

Table of Contents

Chapter 1

Introduction

1.1 Problem of Interest

In statistics, binary logistic regression is used to model the probabilities of a particular pair of (typically mutually exclusive) characteristics. For example, binary logistic regression appears frequently in biostatistical applications as binary responses such as health status (healthy/sick) or survival status (alive/dead) are common. In this thesis, our main concern is solving binary logistic regression problems. Given any data matrix $A \in \mathbb{R}^{N \times n}$ whose rows a_i^T represent input data and response vector $b \in \{-1, 1\}^N$, we assume that

$$P(b_{(i)} = 1 \mid a_i^T; x, y) = \frac{1}{1 + \exp(-a_i^T x + y)}$$

for some parameters $x \in \mathbb{R}^n, y \in \mathbb{R}$ to be estimated. Throughout this thesis, we use $b_{(i)}$ to denote the i-th component of a vector b. Noting that,

$$P(b_{(i)} = -1 \mid a_i^T; x, y) = 1 - P(b_{(i)} = 1 \mid a_i^T; x, y) = \frac{\exp(-a_i^T x + y)}{1 + \exp(-a_i^T x + y)} = \frac{1}{1 + \exp(a_i^T x + y)},$$

we see that the probability mass function can be written as

$$p(b_{(i)} \mid a_i^T; x, y) = \frac{1}{1 + \exp(-b_{(i)} a_i^T x + y)}. \tag{1.1}$$

To estimate parameters x and y, we proceed by maximizing the likelihood function, or equivalently, the log-likelihood function $l : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ given by

$$l(x, y) = \sum_{i=1}^{N} \log\left(\frac{1}{1 + \exp(-b_{(i)} a_i^T x + y)}\right) = -\sum_{i=1}^{N} \log\left(1 + \exp(-b_{(i)} a_i^T x + y)\right).$$

Since a factor of 2 does not change the minimizer, we say that the binary logistic regression problem is an optimization problem of the form

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}} \sum_{i=1}^{N} 2\log\left(1 + \exp\left(-b_{(i)}(a_i^T x + y)\right)\right). \tag{1.2}$$

For theoretical analysis, it may be more helpful to write (1.2) more succinctly. Define for any $u \in \mathbb{R}^k$

$$\begin{aligned} h(u) \equiv h_k(u) &:= \sum_{i=1}^{k} 2\log\left(2\cosh\left(\frac{u_{(i)}}{2}\right)\right) \\ &= \sum_{i=1}^{k} 2\log\left(\exp\left(\frac{u_{(i)}}{2}\right) + \exp\left(-\frac{u_{(i)}}{2}\right)\right). \end{aligned} \tag{1.3}$$

Now, denoting

$$\phi_{A,b}(x, y) := h(Ax + y1) - b^T(Ax + y1) \tag{1.4}$$

where $1_N \in \mathbb{R}^N$ is a vector of 1's, we can reformulate (1.2) as

$$\min_{x\in\mathbb{R}^n, y\in\mathbb{R}} \phi_{A,b}(x, y). \tag{BLR}$$

Indeed, under the definition of h in (1.3), we have

$$\begin{aligned} \phi_{A,b}(x, y) &= \sum_{i=1}^{N} 2\log\left(\exp\left(\frac{a_i^T x + y}{2}\right) + \exp\left(-\frac{a_i^T x + y}{2}\right)\right) - b_{(i)}(a_i^T x + y) \\ &= \sum_{i=1}^{N} 2\log\left(\exp\left(\frac{b_{(i)}(a_i^T x + y)}{2}\right) + \exp\left(-\frac{b_{(i)}(a_i^T x + y)}{2}\right)\right) - b_{(i)}(a_i^T x + y) \\ &= \sum_{i=1}^{N} 2\log\left(1 + \exp\left(-b_{(i)}(a_i^T x + y)\right)\right). \end{aligned}$$

Therefore (BLR) is equivalent to (1.2).

1.2 Properties of Smooth Convex Functions

In this section, we discuss and prove a few properties of smooth convex functions that will be helpful in solving (BLR). In particular, we prove that the log-likelihood function of the binary logistic regression model is convex and smooth. Let us begin with a definition.

Definition 1.1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be convex if for any $x, y \in \mathbb{R}^n$ and any $\lambda \in [0, 1]$,

$$f(\lambda x + (1 - \lambda y)) \leq \lambda f(x) + (1 - \lambda) f(y). \tag{1.5}$$

In addition, we say that f is concave if $-f$ is convex.

When f is differentiable, we have an equivalent definition of convexity:

Proposition 1.1. A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if and only if for any $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle. \tag{1.6}$$

Proof. Let $\lambda \in [0, 1]$ and suppose that f is convex, i.e. (1.5) holds. Then by rearrangement,

$$\begin{aligned} f(y) &\geq \frac{f(\lambda x + (1 - \lambda) y) - \lambda f(x)}{1 - \lambda} \\ &\geq \frac{f(\lambda x + (1 - \lambda) y) - f(x)}{1 - \lambda} + f(x) \end{aligned}$$

holds. Now apply a change of variable $h = 1 - \lambda$ so that we obtain

$$f(x) + \frac{f(1 - h)x + hy) - f(x)}{h} \leq f(y).$$

Letting $h \to 0$, we conclude that

$$\langle \nabla f(x), y - x \rangle + f(x) \leq f(y).$$

Conversely, suppose that (1.6) holds. Then we have

$$\begin{aligned} f(x) &\geq f(\lambda x + (1 - \lambda) y) + \langle \nabla f(\lambda x + (1 - \lambda) y) \rangle, x - \lambda x - (1 - \lambda) y)) \\ f(y) &\geq f(\lambda x + (1 - \lambda) y) + \langle \nabla f(\lambda x + (1 - \lambda) y) \rangle, y - \lambda x - (1 - \lambda) y)). \end{aligned}$$

Let $z = \lambda x + (1-\lambda)y$. Applying these inequalities to $\lambda f(x) + (1-\lambda)y$ gives

$$\begin{aligned}\lambda f(x) + (1-\lambda)f(y) &\geq \lambda\left(f(z) + (1-\lambda)\langle \nabla f(z), x-y\rangle\right) + (1-\lambda)\left(f(z) + \lambda\langle \nabla f(z), y-x\rangle\right)\\ &= f(z).\end{aligned}$$

Thus, f is convex. □

For twice differentiable functions, we may also use the following corollary to verify convexity:

Corollary 1.2. A function $f : \mathbb{R}^n \to \mathbb{R}$ with $\nabla^2 f(x) \succeq 0$ for all $x \in \mathbb{R}^n$ is convex.

Proof. Let f satisfy the corollary and $x, y \in \mathbb{R}^n$. Consider the Taylor series expansion of $f(y)$ about the point x.

$$f(y) = f(x+y-x) = f(x) + (y-x)^T \nabla f(x) + (y-x)^T \frac{\nabla^2 f(\xi)}{2}(y-x)$$

for some $\xi \in \mathbb{R}^n$. Since $\nabla^2 f(\xi) \succeq 0$ by assumption, it follows that

$$f(y) \geq f(x) + (y-x)^T \nabla f(x).$$

Thus f is convex. □

Additionally, there are a few operations that preserve convexity. Here, we list two that will be of use to us.

Proposition 1.3. Let $\{f_i\}_{i=1}^m$ be a set of convex functions with $f_i : \mathbb{R}^n \to \mathbb{R}$ for all i and $\{c_i\}_{i=1}^m$ be a set of constants satisfying $c_i \geq 0$ for all i. Then $f = \sum_{i=1}^m c_i f_i$ is a convex function.

Proof. Let $x, y \in \mathbb{R}^n$ and $\lambda \in [0,1]$. Then we have

$$\begin{aligned}f(\lambda x + (1-\lambda)y) &= \sum_{i=1}^m c_i f_i(\lambda x + (1-\lambda)y)\\ &\geq \sum_{i=1}^m c_i\left(\lambda f_i(x) + (1-\lambda)f_i(y)\right)\\ &= \lambda f(x) + (1-\lambda)f(y).\end{aligned}$$

Thus f is convex. □

Proposition 1.4. Let $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function, $A \in \mathbb{R}^{n\times m}$, and $b \in \mathbb{R}^n$. The function $f : \mathbb{R}^m \to \mathbb{R}$ defined by

$$f(x) = g(Ax + b)$$

is convex.

Proof. Let f, g, A, and b be defined as above, fix points $x, y \in \mathbb{R}^m$, and let $\lambda \in [0, 1]$. Then

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= g(A(\lambda x + (1-\lambda)y) + b) \\ &= g(\lambda(Ax + b) + (1-\lambda)(Ay + b)) \\ &\geq \lambda g(Ax + b) + (1-\lambda)(Ay + b) \\ &= \lambda f(x) + (1-\lambda)f(y). \end{aligned}$$

We conclude that f is convex. □

In addition to convexity, we will also require some smoothness for later sections. The following definition and proposition will prove useful.

Definition 1.2. We say that function $f : \mathbb{R}^n \to \mathbb{R}$ is smooth (with respect to $||\cdot||$) if it is differentiable and ∇f is Lipschitz continuous with Lipschitz constant $L > 0$:

$$||\nabla f(x) - \nabla f(y)||_* \leq L\,||x - y|| \tag{1.7}$$

for any $x, y \in \mathbb{R}^n$. Here, $||\cdot||$ is any norm and $||\cdot||_*$ is its dual norm defined as

$$||x||_* = \sup_{||y||\leq 1} \langle x, y\rangle$$

with $x \in \mathbb{R}^n$. Although this definition holds in general, for the rest of this thesis, we will use the Euclidean norm $||\cdot||_2$. Note that $||\cdot||_2$ is self-dual.

Proposition 1.5. Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function that additionally satisfies (1.7). Then,

$$|f(y) - f(x) - \langle\nabla f(x), y - x\rangle| \leq \frac{L}{2}\,||y - x||^2 \tag{1.8}$$

for all $x, y \in \mathbb{R}^n$.

Proof. We begin by defining an auxillary function

$$g(\lambda) := f((1-\lambda)x + \lambda y) \tag{1.9}$$

for a fixed pair $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$. It follows that by chain rule,

$$g'(\lambda) = \langle \nabla f((1-\lambda)x + \lambda y, y - x\rangle.$$

Since $g(1) = f(y)$ and $g(0) = f(x)$, we may apply the fundamental theorem of calculus to obtain

$$g(1) = g(0) + \int_0^1 g'(\lambda) d\lambda$$

Applying the definition (1.9), we may expand this to

$$f(y) = f(x) + \langle \nabla f(x), y - x\rangle + \int_0^1 \langle \nabla f(1-\lambda)x + \lambda y) - f(x), y - x\rangle d\lambda.$$

By application of Holder's inequality, we obtain

$$\begin{aligned}
|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| &= \left| \int_0^1 \langle \nabla f((1-\lambda)x + \lambda y) - \nabla f(x), y - x\rangle \, d\lambda \right| \\
&\leq \int_0^1 |\langle \nabla f((1-\lambda)x + \lambda y) - \nabla f(x), y - x\rangle| \, d\lambda \\
&\leq \int_0^1 ||\nabla f((1-\lambda)x + \lambda y) - \nabla f(x)||_* \, ||y - x|| \, d\lambda \\
&\leq \int_0^1 L\lambda \, ||y - x||^2 \, d\lambda \\
&= \frac{L}{2} \, ||y - x||^2 .
\end{aligned}$$

□

The following corollary is a direct result of (1.6) and (1.8).

Corollary 1.6. For any convex differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfying (1.7), we have

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - y\rangle \leq \frac{L}{2} \, ||y - x||^2 \tag{1.10}$$

for all $x, y \in \mathbb{R}^n$.

With tools developed in this chapter, we are now ready to show that (BLR) is a smooth convex optimization problem.

Theorem 1.7. The objective function $\phi_{A,b}(x, y)$ in (BLR) is convex and has Lipschitz continuous gradient.

Proof. By Proposition 1.3, to show that function $\phi_{A,b}(x, y) = h(Ax+y1) - b^T(Ax+y1)$ is convex, it suffices to show that $h(Ax+y1)$ and $-b^T(Ax+y1)$ are both convex. The former is the composition of a function h and an affine transformation. By Proposition 1.4, if h is convex, then so is $h(Ax+y1)$. From the definition of h in (1.3), it is easily verified that $\nabla^2 h(x)$ is diagonal with

$$[\nabla^2 h(x)]_{ii} = \frac{\cosh^2\left(\frac{x_{(i)}}{2}\right) + \sinh^2\left(\frac{x_{(i)}}{2}\right)}{\cosh^2\left(\frac{x_{(i)}}{2}\right)} > 0, \forall x.$$

For the latter term, note that $-b^T(Ax + y1)$ is the composition of an affine function and a linear function which is verifiably convex from the definition in (1.5). Thus ϕ_A, b is a convex function. Observe that by defining $z = \begin{pmatrix} x \\ y \end{pmatrix}, D = \begin{pmatrix} A & 1_N \end{pmatrix}$ we have

$$\nabla \phi_{A,b}(z) = D^T \nabla h(Dz) + D^T b.$$

Continuing, let $z_1, z_2 \in \mathbb{R}^{n+1}$ be two vectors. Then

$$\begin{aligned} ||\nabla \phi_{A,b}(z_1) - \nabla \phi_{A,b}(z_2)|| &= \left|\left|D^T \nabla h(Dz) + D^T b - D^T \nabla h(Dz) - D^T b\right|\right| \\ &= \left|\left|D^T \left(\nabla h(Dz_1) - \nabla h(Dz_2)\right)\right|\right| \\ &\leq ||D|| \, ||\nabla h(Dz_1) - \nabla h(Dz_2)|| \, . \end{aligned}$$

Thus, to finish the proof for Lipschitz continuity of $\nabla \phi_{A,b}(x, y)$, it suffices to show that $||\nabla h(Dz_1) - \nabla h(Dz_2)|| \leq L \, ||z_1 - z_2||$ for some constant L. Recalling from the definition of h in (1.3), we see that

$$\nabla h(u) = \tanh\left(\frac{u}{2}\right) := \left(\tanh\left(\frac{u_{(1)}}{2}\right), \ldots, \tanh\left(\frac{u_{(k)}}{2}\right)\right)^T, \forall u \in \mathbb{R}^k, \forall k \qquad (1.11)$$

where we allow the fucntion tanh to be applied component wisely. It follows that

$$||\nabla h(Dz_1) - \nabla h(Dz_2)|| = \left|\left|\tanh(\frac{Dz_1}{2}) - \tanh(\frac{Dz_2}{2})\right|\right|.$$

Since tanh is a Lipschitz continuous function with $L = 1$, we conclude that

$$\left|\left|\tanh(\frac{Dz_1}{2}) - \tanh(\frac{Dz_2}{2})\right|\right| \leq \left|\left|\frac{D}{2}(z_1 - z_2)\right|\right| \leq \frac{||D||}{2} ||z_1 - z_2||$$

and consequently,

$$||\nabla \phi_{A,b}(z_1) - \nabla \phi_{A,b}(z_2)|| \leq \frac{||D||^2}{2} ||z_1 - z_2|| \leq ||D||^2 ||z_1 - z_2|| .$$

Thus, $\nabla \phi_{A,b}$ is Lipschitz continuous with Lipschitz constant $||D||^2$.

□

Chapter 2

Solving the Binary Logistic Regression Problem

Let us now take a step back and tackle a larger class of problems, namely smooth convex optimization. Specifically, consider an optimization problem of the form

$$x^* := \operatorname*{argmin}_{x \in X} f(x) \tag{P}$$

where $X \subset \mathbb{R}^n$ is a convex set and $f : X \to \mathbb{R}$ is a convex function whose gradient satisfies (1.7) with Lipschitz constant $L > 0$. In later sections of this chapter, we will discuss methods for solving (P) along with convergence analysis and limitations of such a scheme. Recall that in Theorem 1.7, we showed that the function $\phi_{A,b}$ defined in (1.4) satisfies these conditions. Thus, by solving (P), we will also develop a method of solving (BLR). Here, we provide an algorithm capable of computing an ε-approximate solution $\hat{x}$ such that $f(\hat{x}) - f^* \leq \varepsilon$ in $\mathcal{O}(1)(1/\sqrt{\varepsilon})$ iterations based on the original work presented in [1].

2.1 A First Order Method for Smooth Convex Optimization

We now present Nesterov's accelerated gradient descent for solving (P).

Algorithm 1 Nesterov's accelerated gradient descent (NAGD)

```
Select parameters $\gamma_k \in (0,1], \eta_k$. Choose $x_0 \in X$. Set $y_0 = x_0$.
for $k = 1, \dots, N$ do
    $z_k = (1-\gamma_k) y_{k-1} + \gamma_k x_{k-1}$    (2.1)
    $x_k = \underset{x \in X}{\mathrm{argmin}} \langle \nabla f(z_k), x \rangle + \frac{\eta_k}{2} ||x_{k-1} - x||_2^2$    (2.2)
    $y_k = (1-\gamma_k) y_{k-1} + \gamma_k x_k$    (2.3)
end for
Output $y_N$.
```

Note that in order to use Algorithm 1 to solve (P), we only require function and gradient evaluations, an initial feasible point, and constants γ_k and η_k. We will see in the sequel that, for different choices of γ_k, η_k, Algorithm 1 may have different performances.

2.2 Convergence Analysis of Accelerated Gradient Descent

In the following theorem, we state the convergence result of the accelerated gradient descent in Algorithm 1.

Theorem 2.1. Suppose that we apply Algorithm 1 to solve (P) with parameter $\gamma_k \in [0,1]$. Then the k-th iterates satisfy

$$f(y_k) - (1-\gamma_k) f(y_{k-1}) - \gamma_k f(x) \leq \gamma_k \eta_k \left(||x - x_{k-1}||^2 - ||x - x_k||^2 \right) + \frac{L\gamma_k^2 - \eta_k \gamma_k}{2} ||x_k - x_{k-1}||^2 . \tag{2.4}$$

Proof. Let $\{z_i\}_{i=0}^k, \{x_i\}_{i=0}^k$ and $\{y_i\}_{i=0}^k$ be the iterate sequences generated by Algorithm 1. Then by Lipschitz continuity of gradients and convexity, we have the following:

$$f(y_k) \leq f(z_k) + \langle \nabla f(z_k), y_k - z_k \rangle + \frac{L}{2} ||y_k - z_k||^2 \tag{2.5}$$

$$f(y_{k-1}) \geq f(z_k) + \langle \nabla f(z_k), y_{k-1} - z_k \rangle \tag{2.6}$$

$$f(x) \geq f(z_k) + \langle \nabla f(z_k), x - z_k \rangle . \tag{2.7}$$

Each inequality (2.5), (2.6), and (2.7) follows immediately from (1.8). With these inequalities in hand, we write

$$\begin{aligned}f(y_k) - (1-\gamma_k)f(y_{k-1}) - \gamma_k f(x) \leq{}& f(z_k) + \langle \nabla f(z_k), y_k - z_k\rangle + \frac{L}{2}\left\|y_k - z_k\right\|^2\\ &- (1-\gamma_k)(f(z_k) + \langle \nabla f(z_k), y_{k-1} - z_k\rangle)\\ &- \gamma_k(f(z_k) + \langle \nabla f(z_k), x - z_k\rangle).\end{aligned}$$

After simplifying and noting $y_k - z_k = \gamma_k(x_k - x_{k-1})$ from (2.3) and (2.1), we have

$$\begin{aligned}f(y_k) - (1-\gamma_k)f(y_{k-1}) - \gamma_k f(x) &\leq \langle \nabla f(z_k), y_k - (1-\gamma_k)y_{k-1} - \gamma_k x\rangle + \frac{L\gamma_k^2}{2}\left\|x_k - x_{k-1}\right\|^2\\ &= \langle \nabla f(z_k), y_k - (1-\gamma_k)y_{k-1} - \gamma_k x_k + \gamma_k x_k - \gamma_k x\rangle + \frac{L\gamma_k^2}{2}\left\|x_k - x_{k-1}\right\|^2\\ &= \gamma_k\langle \nabla f(z_k), x_k - x\rangle + \frac{L\gamma_k^2}{2}\left\|x_k - x_{k-1}\right\|^2.\end{aligned}$$

Here, we make use of the definition of y_k in (2.3) in the final step. Enforcing the optimality conditions of x_k in (2.2) to $\gamma_k\langle \nabla f(z_k), x_k - x\rangle$, we obtain

$$f(y_k) - (1-\gamma_k)f(y_{k-1}) - \gamma_k f(x) \leq \gamma_k\eta_k\left(\left\|x - x_{k-1}\right\|^2 - \left\|x - x_k\right\|^2\right) + \frac{L\gamma_k^2 - \eta_k\gamma_k}{2}\left\|x_k - x_{k-1}\right\|^2$$

which completes our proof. □

With Theorem 2.1 in hand, we state two corollaries specifying choices of parameters γ_k and η_k.

Corollary 2.2. If we set

$$\gamma_k \equiv 1 \text{ and } \eta_k \equiv L \tag{2.8}$$

in Algorithm 1, then

$$f(\tilde{y}_N) - f(x^*) \leq \frac{L\left\|x^* - x_0\right\|^2}{N+1} \tag{2.9}$$

where $\tilde{y}_N = \sum_{k=0}^N y_k/(N+1)$.

Proof. Note that by convexity of f,

$$f(\tilde{y}_N) = f\left(\frac{\sum_{k=0}^N y_k}{N+1}\right) \leq \frac{1}{N+1}\sum_{k=0}^N f(y_k). \tag{2.10}$$

To prove the corollary, it suffices to show that

$$\frac{1}{N+1}\left(\sum_{k=0}^{N} f(y_k)\right) - f(x^*) \leq \frac{L\,||x^* - x_0||^2}{N+1}.$$

Following the result of Theorem 2.1 with $\gamma_k \equiv 1$ and $\eta_k \equiv L$, (2.4) becomes

$$f(y_k) - f(x) \leq L\left(||x - x_{k-1}||^2 - ||x - x_k||^2\right) \tag{2.11}$$

Summing (2.11) over all k, the right hand side becomes a telescoping sum to yield

$$\left(\sum_{k=0}^{N} f(y_k)\right) - (N+1)f(x) \leq L\left(||x - x_0||^2 - ||x - x_N||^2\right) \leq L\,||x - x_0||^2.$$

Setting $x = x^*$ and dividing by $(N+1)$ gives the desired result. □

The parameter setting in (2.8) is desirable in part due to its simplicity. Indeed, under these settings, $x_k = y_k = z_k$ and

$$\begin{aligned} x_k &= \operatorname*{argmin}_{x \in X} \langle \nabla f(x_k), x\rangle + \frac{L}{2}\,||x_{k-1} - x||^2 \\ &= \operatorname*{argmin}_{x \in X} \frac{L}{2}\left|\left|x - \left(x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})\right)\right|\right|^2 \\ &= \operatorname*{argmin}_{x \in X} \left|\left|x - \left(x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})\right)\right|\right|^2 \end{aligned}$$

We may understand the above computation of x_k as a step in the negative direction of $\nabla f(x_{k-1})$ projected back into the set X. This technique is referred to as Projected Gradient Descent and is presented in Algorithm 2.

Algorithm 2 Projected Gradient Descent

```
Choose $x_0 \in X$.
for $k = 1, \ldots, N$ do
    $x_k = \operatorname*{argmin}_{x \in X} \left|\left|x - (x_{k-1} - \frac{1}{L}\nabla f(x_{k-1})\right|\right|^2$
end for
Output $\tilde{x}_N = \sum_{k=0}^{N} x_k/(N+1)$.
```

Corollary 2.3. If we set

$$\gamma_k = \frac{2}{k+1} \text{ and } \eta_k = \frac{2L}{k} \tag{2.12}$$

in Algorithm 1, then

$$f(y_N) - f(x^*) \leq \frac{4L}{N(N+1)} \left\|x^* - x_0\right\|^2. \tag{2.13}$$

Proof. Define

$$\Gamma_k = \begin{cases} 1 & k = 1 \\ (1-\gamma_k)\Gamma_{k-1} & k > 1 \end{cases} \tag{2.14}$$

Under the setting in (2.12), it is easy to verify that $\Gamma_k = \frac{2}{k(k+1)}$. Letting $x = x^*$ in (2.4) and dividing by Γ_k gives

$$\frac{1}{\Gamma_k}(f(y_k) - f(x^*)) \leq \frac{1-\gamma_k}{\Gamma_k}(f(y_{k-1}) - f(x^*)) + \frac{\gamma_k \eta_k}{\Gamma_k}\left(\left\|x^* - x_{k-1}\right\|^2 - \left\|x_k - x^*\right\|^2\right).$$

Subtracting the first term on the right hand side from each side and summing over k gives us another telescoping series on the left that evaluates to

$$f(y_1) - f(x^*) + \sum_{k=2}^{N} \frac{1}{\Gamma_k}(f(y_k) - f(x^*)) - \frac{1}{\Gamma_{k-1}}(f(y_{k-1}) - f(x^*)) \leq \sum_{k=1}^{N} \frac{\gamma_k \eta_k}{\Gamma_k}(\left\|x^* - x_{k-1}\right\|^2 - \left\|x_k - x^*\right\|^2) \tag{2.15}$$

since $(1-\gamma_1)/\Gamma_1 = 0$. Evaluating the telescoping series on the, we obtain

$$\frac{1}{\Gamma_N}(f(y_N) - f(x^*)) \leq \sum_{k=1}^{N} \frac{\gamma_k \eta_k}{\Gamma_k}(\left\|x^* - x_{k-1}\right\|^2 - \left\|x_k - x^*\right\|^2). \tag{2.16}$$

Applying the parameters (2.12) and (2.14) gives

$$\frac{N(N+1)}{2}(f(y_N) - f(x^*)) \leq \sum_{k=1}^{N} \frac{4Lk(k+1)}{2k(k+1)}(\left\|x^* - x_{k-1}\right\|^2 - \left\|x_k - x^*\right\|^2)$$

that simplifies to

$$f(y_N) - f(x^*) \leq \frac{4L}{N(N+1)} \left\|x^* - x_0\right\|^2.$$

□

Recall that from (2.9), in order to compute an approximate solution y such that $f(y) -$

$f(x^*) \leq \varepsilon$ with parameters defined in (2.8), the number of iterations required is bounded by $\mathcal{O}(\frac{L||x^*-x_0||^2}{\varepsilon})$. However, under (2.12), (2.13) shows that number is bounded by $\mathcal{O}\left(\sqrt{\frac{4L||x^*-x_0||^2}{\varepsilon}}\right)$. While the former reduces to the simple scheme in Algorithm 2, the latter gives accelerated convergence and is commonly referred to as Nesterov's optimal gradient method (OGD).

Chapter 3

Optimality of Nesterov's Optimal Gradient Method

Algorithm 1 is an example of a first order iterative method. We define a first order method $\mathcal{M}$ as a scheme for solving (P) such that $\mathcal{M}$ initializes the search point x_0, and $\mathcal{M}$ accesses the first order information of f through a deterministic oracle $\mathcal{O}_f : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^n$ with $\mathcal{O}_f(x) = (f(x), \nabla f(x))$ for $x \in \mathbb{R}^n$. In particular, $\mathcal{M}$ can be described by a problem independent x_0 and a sequence of rules $\{\mathcal{I}\}_{k=0}^{\infty}$ such that

$$x_{k+1} = \mathcal{I}(\mathcal{O}_f(x_0), \ldots, \mathcal{O}_f(x_k)).$$

Without loss of generality, we may assume that $x_0 = 0$ and that at the N-th iteration, the output of $\mathcal{M}$ is always x_N.

In [1], it is shown that the lower complexity bound for smooth convex optimization is $\mathcal{O}(1)(1/\sqrt{\varepsilon})$. In view of Corollary 2.3, we conclude that OGD is an optimal algorithm for smooth convex optimization, in the sense that its theoretical computational performance can not be improved. For completeness, we present a proof in this chapter.

3.1 Iteration Complexity Lower Bound for Smooth Convex Optimization

We will consider a quadratic programming problem, which falls under the class of smooth convex optimization, of the form

$$f^* := \min_{x \in \mathbb{R}^n} f(x) := Q_{A,b}(x) := \frac{1}{2} x^T A x - b^T x \tag{QP}$$

where $A \succeq 0$. Convexity of (QP) follows immediately from Corollary (1.2) since $\nabla^2 f(x) = A \succeq 0$ by assumption. Furthermore, for any $U \in \mathcal{U}_b$ where

$$\mathcal{U}_b := \{U \in \mathbb{R}^{n \times n} \mid Ub = U^T b = b \text{ and } U \text{ is orthogonal}\} \tag{3.1}$$

we can show that by defining $y = Ux$ for any $x \in \mathbb{R}^n$

$$x^T U^T A U x = y^T A y \geq 0$$

follows from $A \succeq 0$. Thus,

$$f_U(x) := f(Ux) = \frac{1}{2} x^T U^T A U x - b^T x = Q_{U^T AU, b}(x) \tag{3.2}$$

is also a smooth convex function and

$$\min_{x \in \mathbb{R}^n} f(x) = \min_{x \in \mathbb{R}^n} f_U(x).$$

That is, without loss of generality, we may assume our quadratic program has objective function $f_U(x)$ for some $U \in \mathcal{U}_b$. Let $\mathcal{K}_r(A, b)$ be the Krylov subspace of order r defined by

$$\mathcal{K}_r(A, b) := \operatorname{span}\{b, Ab, \ldots, A^{r-1} b\}. \tag{3.3}$$

We have the following lemmas.

Lemma 3.1. For any $A \in \mathbb{R}^{n \times n}, b \in \mathbb{R}^n$, we have the following

1. $\mathcal{K}_r(A,b) \subseteq \mathcal{K}_{r+1}(A,b)$ for any integer $r \geq 1$. Additionally, if $\mathcal{K}_r(A,b) = \mathcal{K}_{r+1}(A,b)$, then $\mathcal{K}_r(A,b) = \mathcal{K}_s(A,b)$ for all integers $s \geq r$.

2. For all $U \in \mathcal{U}_b$,
$$\mathcal{K}_r(U^T AU, b) = U^T \mathcal{K}_r(A,b). \tag{3.4}$$

Proof. The subset $\mathcal{K}_r(A,b) \subseteq \mathcal{K}_{r+1}(A,b)$ is clear from the definition in (3.3). For the remainder of the first part, assume that $\mathcal{K}_r(A,b) = \mathcal{K}_{r+1}(A,b)$. That is, there must exist $c_i \in \mathbb{R}$ such that

$$A^r b = \sum_{i=1}^{r} c_i A^{i-1} b$$

This implies that

$$A^{r+1}b = A(A^r b) = A \sum_{i=1}^{r} c_i A^{i-1} b = \sum_{i=1}^{r} c_i A^i b = \sum_{i=1}^{r-1} c_i A^i b + c_r A^r b.$$

The term $\sum_{i=1}^{r-1} c_i A^i b$ lies in $\mathcal{K}_r(A,b)$ by definition whereas $c_r A^r b \in \mathcal{K}_{r+1}(A,b) = \mathcal{K}_r(A,b)$ by assumption. Thus, $A^{r+1}b \in \mathcal{K}_r(A,b)$. Inducting on r gives the desired result. For the second statement, note from (3.1) that

$$(U^T AU)^i b = U^T A^i U b = U^T A^i b.$$

Thus, $x = \sum_{i=1}^{r} c_i (U^T AU)^{i-1} b = U^T \sum_{i=1}^{r} c_i A^{i-1} b$ which implies that $\mathcal{K}_r(U^T AU, b) = U^T \mathcal{K}_r(A,b)$. □

The following lemma appeared in [2] and will be used in our analysis in the sequel.

Lemma 3.2. Let X and Y be two linear subspaces satisfying $X \subsetneq Y \subseteq \mathbb{R}^p$. Then for any $\bar{x} \in \mathbb{R}^p$, there exists orthogonal matrix V such that

$$V\bar{x} \in Y \text{ and } Vx = x, \, \forall x \in X$$

Proof. When $\bar{x} \in X$, $V = I$ satisfies the conditions. Otherwise, let $\bar{x} = y + z$ where $z \in X, y \in X^{\perp}$. Let $\{u_i\}_{i=1}^{t}, t < p$ be an orthonormal basis for X and extend it to an orthonormal basis $\{u_i\}_{i=1}^{s}, t <$

$s \leq p$ for Y. Define an orthogonal matrix V such that

$$Vu_i = u_i \text{ and } Vy = ||y||\, u_{t+1}.$$

Then for any $x \in X$, there exists constants λ_i such that $\sum_{i=1}^{t} \lambda_i u_i = x$ and consequently

$$Vx = \sum_{i=1}^{t} \lambda_i V u_i = \sum_{i=1}^{t} \lambda_i u_i = x.$$

Furthermore, it follows directly from construction that

$$V\bar{x} = z + ||y||\, u_{t+1} \in Y.$$

□

The following two propositions will be crucial to development of the lower complexity bound for smooth convex optimization.

Proposition 3.1. Let $f(x)$ be defined as in (QP) and N an integer such that

$$\mathcal{K}_{2N-1}(A, b) \subsetneq \mathcal{K}_{2N}(A, b) \subsetneq \mathcal{K}_{2N+1}(A, b). \tag{3.5}$$

Then for any first order scheme $\mathcal{M}$ and iterate number k, there exists $U_k \in \mathcal{U}_b$ such that when $\mathcal{M}$ is applied to minimize $f_{U_k}(x)$, the first k iterates satisfy

$$x_i \in U_k^T \mathcal{K}_{2k+1}(A, b)$$

for all $i = 0, \ldots, k$.

Proof. We proceed by induction. When $k = 0$, the result is trivial since for any $U_0 \in \mathcal{U}_b$, we have $x_0 = 0 \in U_0 \operatorname{span}\{b\}$. Now assume the statement holds for $k - 1 < N$ and let x_k be the next iterate computed by $\mathcal{M}$. By the assumption $k - 1 < N$ and (3.5),

$$\mathcal{K}_{2k-1}(A, b) \subsetneq \mathcal{K}_{2k}(A, b) \subsetneq \mathcal{K}_{2k+1}(A, b).$$

and consequently

$$U_{k-1}^T\mathcal{K}_{2k-1}(A,b) \subsetneq U_{k-1}^T\mathcal{K}_{2k}(A,b) \subsetneq U_{k-1}^T\mathcal{K}_{2k+1}(A,b).$$

by orthogonality of U_{k-1}. Suppose that $U_{k-1}^T\mathcal{K}_{2k}(A,b)$ is spanned by $\{w_1,\ldots,w_l\}$ for some $l \leq 2k$. Then there must exist $w_{l+1} \in U_{k-1}^T\mathcal{K}_{2k+1}(A,b)$ for which $w_{l+1} \notin U_{k-1}^T\mathcal{K}_{2k}(A,b)$. By Lemma 3.2 we can define an orthogonal matrix V satisfying

$$Vw_i = w_i \text{ for } w_i \in U_{k-1}^T\mathcal{K}_{2k}(A,b) \text{ and } Vv_k/\left\|v_k\right\| = w_{l+1}. \tag{3.6}$$

By this definition, it follows that $x_k \in V^TU_{k-1}^T\mathcal{K}_{2k+1}(A,b)$. We claim that $U_k := U_{k-1}V$ is the orthogonal matrix we seek to complete the induction step. It is easy to see that U_k is orthogonal and since $b \in U_{k-1}^T\mathcal{K}_{2k}(A,b)$,

$$U_kb = U_{k-1}Vb = U_{k-1}b = b.$$

It suffices to show that applying $\mathcal{M}$ to $f_{U_k}(x)$ generates iterates $x_0,\ldots,x_{k-1}$. Indeed, note that for any $x \in U_{k-1}^T\mathcal{K}_{2k-1}(A,b)$, from (3.4) we have that $U_{k-1}^TAU_{k-1}x \in U_{k-1}^T\mathcal{K}_{2k}(A,b)$. Thus

$$U_k^TAU_kx = V^TU_{k-1}^TAU_{k-1}Vx = V^T(U_{k-1}^TAU_{k-1}x) = U_{k-1}^TAU_{k-1}x.$$

Hence, for any $x \in U_{k-1}^T\mathcal{K}_{2k-1}(A,b)$, the functions $f_{U_k}(x)$ and $f_{U_{k-1}}(x)$ have the same zeroth and first order information. Since we assume by the induction hypothesis that $x_0,\ldots,x_{k-1} \in U_{k-1}^T\mathcal{K}_{2k-1}(A,b)$, applying $\mathcal{M}$ to minimize $f_{U_k}(x)$ will generate $x_0,\ldots,x_k$. Noting (3.6), we conclude that $x_0,\ldots,x_k \in U_k^T\mathcal{K}_{2k+1}(A,b)$. □

Proposition 3.2. For any $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n, U \in \mathcal{U}_b$, and any integer r,

$$\min_{x\in\mathcal{K}_r(A,b)} Q_{A,b}(x) - \min_{x\in\mathbb{R}^n} Q_{A,b}(x) = \min_{x\in\mathcal{K}_r(U^TAUb)} Q_{U^TAU,b}(x) - \min_{x\in\mathbb{R}^n} Q_{U^TAU,b}(x). \tag{3.7}$$

Proof. Since $U \in \mathcal{U}_b$, $b^T x = U^T U b^T x = U^T b^T x = b^T(Ux)$. Noting (3.4),

$$\begin{aligned}\min_{x \in \mathcal{K}_r(A,b)} \frac{1}{2} x^T U^T A U x - b^T x &= \min_{x \in \mathcal{K}_r(A,b)} \frac{1}{2}(x^T U^T) A (Ux) - b^T U x \\ &= \min_{x \in U^T \mathcal{K}_r(A,b)} \frac{1}{2} x^T A x - b^T x \\ &= \min_{x \in \mathcal{K}_r(U^T A U, b)} \frac{1}{2} x^T A x - b^T x.\end{aligned}$$

Also,

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T U^T A U x - b^T x = \min_{x \in \mathbb{R}^n} \frac{1}{2} x^T U^T A U x - b^T U x = \min_{y \in \mathbb{R}^n} \frac{1}{2} y^T A y - b^T y$$

with $y = Ax$. Combining these equalities gives the result. □

In view of (3.7), we may simply consider an instance of A and b. Indeed, let

$$A = \frac{L}{4} \begin{pmatrix} A_{4k+3} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n \times n}, b = \frac{L}{4} e_1 \tag{3.8}$$

with A_{4k+3} a tridiagonal matrix defined as

$$A_{4k+3} = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

In this definition, L is the Lipschitz constant and e_i is the i-th standard basis vector of $\mathbb{R}^n$. We remark that from [3], the eigenvalues of A are of the form $\lambda_j = 2 - 2\cos(j\pi/(4k+4)) \leq 4$ for all $j = 1, \ldots, 4k+3$. Since A is symmetric, it follows that

$$||A||_2 = \frac{L}{4} \, ||A_{4k+3}|| = \frac{L}{4} \max_i |\lambda_j| \leq \frac{4L}{4} = L.$$

Thus, $\nabla Q_{A,b}(x) = Ax - b$ is Lipschitz continuous with Lipschitz constant L.

The optimal solution x^* to $\min\limits_{x \in \mathbb{R}^n} Q_{A,b}(x)$ is simply the solution to $Ax = b$ and can be verified to be

$$x_i^* = \begin{cases} 1 - \frac{i}{4k+4} & 1 \leq i \leq 4k+3 \\ 0 & 4k+4 \leq i \leq n \end{cases}.$$

and so,

$$Q_{A,b}(x^*) = \min_{x \in \mathbb{R}^n} Q_{A,b}(x) = \frac{L}{8}\left(-1 + \frac{1}{4k+4}\right). \tag{3.9}$$

In the lemma that follows, we show the final necessary computation for proving the lower complexity bound.

Lemma 3.3. For A, b defined in (3.8), the Krylov subspaces can be written as

$$\mathcal{K}_r(A,b) = \begin{cases} \text{span}\{e_1, \ldots, e_r\} & 1 \leq r \leq 4k+2 \\ \text{span}\{e_1, \ldots, e_{4k+3}\} & r \geq 4k+3, \end{cases} \tag{3.10}$$

the minimum value over $\mathcal{K}_{2k+1}(A,b)$ satisfies

$$\min_{x \in \mathcal{K}_{2k+1}(A,b)} Q_{A,b}(x) \geq \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right),$$

and consequently,

$$\min_{x \in \mathcal{K}_{2k+1}(A,b)} Q_{A,b}(x) - \min_{x \in \mathbb{R}^n} Q_{A,b}(x) \geq \frac{3L\,||x^*||^2}{128(k+1)^2}. \tag{3.11}$$

Proof. Clearly, $b \in \text{span}\{e_1\}$ and $Ae_1 = (2, -1, 0, \ldots, 0)^T \in \text{span}\{e_1, e_2\}$. Continuing, we have $Ae_i \in \text{span}\{e_{i-1}, e_i, e_{i+1}\}$ for all $2 \leq i \leq 4k+2$ from the tridiagonal structure. Thus, the first statement follows. Since $x_{(i)} = 0$ for any $i \geq 2k+2$ whenever $x \in \mathcal{K}_{2k+1}(A,b)$, it follows that the error can be bounded as

$$\min_{x \in \mathcal{K}_{2k+1}(A,b)} Q_{A,b}(x) = \min_{x \in \mathcal{K}_{2k+1}(A,b)} \frac{L}{4}\left(\frac{1}{2}x^T A x - b^T x\right) \geq \min_{z \in \mathbb{R}^{2k+1}} \frac{L}{4}\left(\frac{1}{2}z^T A_{2k+1} z - z_{(1)}\right).$$

Here, A_{2k+1} is the $(2k+1) \times (2k+1)$ leftmost submatrix of A and z is the subvector consisting of the first $2k+1$ elements of x. In a similar computation that produced $\min\limits_{x \in \mathbb{R}^n} Q_{A,b}(x)$, we see that

$$\min_{x \in \mathcal{K}_{2k+1}(A,b)} Q_{A,b}(x) \geq \frac{L}{8}\left(-1 + \frac{1}{2k+2}\right).$$

Consequently, the error in terms of the objective function value satisfies

$$\min_{x\in\mathcal{K}_{2k+1}(A,b)} Q_{A,b}(x) - \min_{x\in\mathbb{R}^n} Q_{A,b}(x) \geq \frac{L}{32(k+1)}. \tag{3.12}$$

Combining (3.12) with

$$\begin{aligned}||x^*||^2 &= \sum_{i=1}^{4k+3}\left(1-\frac{i}{4k+4}\right)^2 = \left(\sum_{i=1}^{4k+3} 1\right) - \left(\frac{2}{4k+4}\sum_{i=1}^{4k+3} i\right) + \left(\frac{1}{(4k+4)^2}\sum_{i=1}^{4k+3} i^2\right)\\ &\leq (4k+3) - \frac{2}{4k+4}\cdot\frac{(4k+3)(4k+4)}{2} + \frac{1}{(4k+4)^2}\cdot\frac{(4k+4)^3}{3}\\ &= \frac{4(k+1)}{3}\end{aligned}$$

we conclude (3.11). □

Theorem 3.3. For any first order iterative method $\mathcal{M}$ and iterate $k \leq \frac{n-3}{4}$, there exists some smooth convex function $g : \mathbb{R}^n \to \mathbb{R}$ with L-Lipschitz gradient such that x_k generated by $\mathcal{M}$ satisfies

$$h(x_k) - \min_{x\in\mathbb{R}^n} h(x) \geq \frac{3L\,||x_0 - x^*||^2}{128(k+1)^2}. \tag{3.13}$$

Proof. Let $f : \mathbb{R}^n \to \mathbb{R}$ be defined as in (QP) with A and b as stated in (3.8). Applying Proposition 3.1, there exists $g(x) := f_{U_k}(x)$ whose iterates $x_0, \ldots, x_k$ lies in the subspace $U_k^T\mathcal{K}_{2k+1}(A,b)$. Thus, by Proposition 3.2 and Lemma 3.3,

$$\begin{aligned}\min_{x\in\mathcal{K}_r(U^TAUb)} Q_{U^TAU,b}(x) - \min_{x\in\mathbb{R}^n} Q_{U^TAU,b}(x) &= \min_{x\in\mathcal{K}_r(A,b)} Q_{A,b}(x) - \min_{x\in\mathbb{R}^n} Q_{A,b}(x)\\ &\geq \frac{3L\,||x^*||^2}{128(k+1)^2}.\end{aligned}$$

□

In view of Theorem 3.3, our proposed Algorithm 1 with parameters (2.12) is an optimal algorithm for solving smooth convex optimization. Futhermore, there exists the following other available lower complexity bound results on deterministic first order methods for convex optimization $f^* := \min_x f(x)$.

- When f is convex, the lower complexity bound is $\mathcal{O}(1)(1/\varepsilon^2)$ [4, 1].

- When f is convex, nonsmooth with bilinear saddle point structure, the lower complexity bound is $\mathcal{O}(1)(1/\varepsilon)$ [2].
- When f is strongly convex, smooth the lower complexity bound is $\mathcal{O}(1)\log(1/\varepsilon)$ [1, 5].

Chapter 4

Lower Complexity Bound for Binary Logistic Regression

Let us now consider our (BLR) again. By (1.7), we know that we can solve (BLR) via Algorithm 1. Hence, in view of (2.3), an $\varepsilon-$ approoximate solution can be computed in $\mathcal{O}(1)(1/\sqrt{\varepsilon})$ first order oracle iterations. However, it has yet to be determined that that Algorithm 1 achieves the lower complexity bound for binary logistic regression problems via first order deterministic methods. We will develop a lower complexity bound for functions of the form in (BLR), a subset of smooth convex optimization. In doing so, we will construct a worst-case dataset for solving binary logistic regression that requires $\mathcal{O}(1)(1/\sqrt{\varepsilon})$ first order oracle calls and thus, showing Algorithm 1 is an optimal algorithm for this class of problems. These worst-case constructions will satisfy $y^* = 0$. Consequently, it suffices to solve the logistic model with homogeneous linear predictor

$$l_{A,b}(x) = h(Ax) - b^T Ax \tag{4.1}$$

and corresponding problem

$$l^*_{A,b} = \min_{x\in\mathbb{R}^n} l_{A,b}(x). \tag{HBLR}$$

4.1 Linear Span Assumption

In this section, we make the following simplifying assumption: the iterates of a first order deterministic method $\mathcal{M}$ satisfy

$$x_t \in \operatorname{span}\{\nabla f(x_0), \ldots, \nabla f(x_{t-1})\} \tag{4.2}$$

for all $t \geq 1$. We will refer to (4.2) as the linear span assumption. This assumption is convenient in demonstrating our desired result, but we will later show that our results can also be proved without the linear span assumption through a technique developed in [6] (see also [2]).

Define the following for a parameterization of binary logistic regression

$$W_k := \begin{pmatrix} & & & -1 & 1 \\ & & -1 & 1 & \\ & \iddots & \iddots & & \\ -1 & 1 & & & \\ -1 & & & & \end{pmatrix} \in \mathbb{R}^{k \times k}, A_k := \begin{pmatrix} 2\sigma W_k \\ -2\zeta W_k \\ -2\sigma W_k \\ 2\zeta W_k \end{pmatrix} \in \mathbb{R}^{4k \times k}, b_k = \begin{pmatrix} 1_{2k} \\ -1_{2k} \end{pmatrix} \in \mathbb{R}^{4k}. \tag{4.3}$$

Note that the above construction follows the idea in [1]; indeed, W_k^2 appeared in the construction (3.8). Let us consider performing binary logistic regression with data matrix A_k and response vector b_k. In order to avoid duplicate data entries in A_k, we also assume without loss of generality that $\sigma > \zeta > 0$. Denote then the functions

$$f_k(x) := h(A_kx) - b_k^T(A_kx) \text{ and } \phi_k(x, y) := h(A_kx_{+}y1_k) - b_k^T(A_kx + y1_k). \tag{4.4}$$

In view of the discussion above, we can view f_k and ϕ_k as objective functions of homoegeneous and inhomogeneous binary logistic regression problems respectively. Before we proceed, we note that a sufficient optimality condition for minimizing the function f in (4.4) is

$$\nabla f_k(x) = 0. \tag{4.5}$$

We will use three lemmas to study the behavior of the iterates of a first order method $\mathcal{M}$ applied to (HBLR). For brevity, denote $e_{t,k}$ the t-th standard basis in $\mathbb{R}^k$.

Lemma 4.1. For A_k defined in (3.8),

$$||A_k|| \leq \sqrt{32(\sigma^2 + \zeta^2)}. \tag{4.6}$$

Proof. Let $u \in \mathbb{R}^k$. Then we have

$$\begin{aligned} ||W_k u||^2 &= (u_{(k)} - u_{(k-1)})^2 + \cdots + (u_{(2)} - u_{(1)})^2 + (u_{(1)})^2 \\ &\leq 2\left((u_{(k)})^2 + (u_{(k-1)})^2 + \cdots + (u_{(2)})^2 + (u_{(1)})^2 + (u_{(1)})^2\right) \\ &\leq 4\,||u||^2 . \end{aligned}$$

This implies that

$$||A_k u||^2 = 8(\sigma^2 + \zeta^2)\,||W_k u||^2 \leq 32(\sigma^2 + \zeta^2)\,||u||^2 .$$

Thus,

$$||A_k|| \leq \sqrt{32(\sigma^2 + \zeta^2)}.$$

□

Lemma 4.2. The minimization problem

$$f_k^* := \min_{x \in \mathbb{R}^k} f_k$$

has optimal solution

$$x^* = c(1, 2, \ldots, k)^T$$

with optimal value

$$f_k^* = 8k \log 2 + 4k\,(\log\cosh(\sigma c) + \log\cosh(\zeta c) - (\sigma - \zeta)c)$$

where c satisfies

$$\sigma \tanh(\sigma c) + \zeta \tanh(\zeta c) = \sigma - \zeta. \tag{4.7}$$

In addition, $(x^*, 0)$ is the unique optimal solution to $\min_{x \in \mathbb{R}^n, y \in \mathbb{R}} \phi_k(x, y)$.

Proof. Since $W_k x^* = c1_k$ follows from the definition in (3.8), it is then easy to check that

$$\nabla h(A_k x^*) = \tanh\left(\frac{1}{2}\begin{pmatrix} 2\sigma c 1_k \\ -2\zeta c 1_k \\ -2\sigma c 1_k \\ 2\zeta c 1_k \end{pmatrix}\right) = \begin{pmatrix} \tanh(\sigma c) 1_k \\ \tanh(\sigma c) 1_k \\ \tanh(\sigma c) 1_k \\ \tanh(\sigma c) 1_k \end{pmatrix}$$

via the description in (1.11). Using above equation and noting that $W_k 1_k = e_{k,k}$, we have

$$A_k^T \nabla h(A_k x^*) = 4(\sigma \tanh(\sigma c) + \zeta \tanh(\zeta c)) e_{k,k}.$$

Combining the above with

$$A_k^T b_k = 4(\sigma - \zeta) e_{k,k} \tag{4.8}$$

we conclude that the optimality condition in (4.5) is met whenever c satisfies (4.7). Such a c is well-defined since the function $T(c) := \sigma \tanh(\sigma c) + \zeta \tanh(\zeta c) - \sigma + \zeta$ is continuous with $T(0) = -\sigma + \zeta < 0$ and $\lim_{c \to \infty} T(c) = 2\zeta > 0$. Moreover, noting that cosh is an even function,

$$\begin{aligned} f_k^* &= h(A_k x^*) - 4k(\sigma - \zeta)c \\ &= 2k\left(\log(2\cosh(\sigma c)) + \log(2\cosh(-\zeta c)) + \log(2\cosh(\sigma c)) + \log(2\cosh(\zeta c))\right) - 4k(\sigma - \zeta)c \\ &= 8k \log 2 + 4k\left(\log\cosh(\sigma c) + \log\cosh(\zeta c) - (\sigma - \zeta)c\right). \end{aligned}$$

Furthermore, after noting

$$\left.\frac{\partial}{\partial y}\right|_{y=0} \phi_k(x^*, y) = 1_k^T \nabla h(A_k x^*) - b_k^T 1_k = 0$$

and

$$\nabla_x \phi_k(x^*, 0) = \nabla f_k(x^*) = 0,$$

we conclude that

□

Lemma 4.3. Let t, k be positive integers satisfying $t \leq k$. Denote

$$\mathcal{X}_{t,k} := \operatorname{span}\{e_{k-t+1,k}, \ldots, e_{k,k}\},\ \forall k,\ 1 \leq t \leq k$$

and

$$\mathcal{Y}_{t,k} := \operatorname{span}\{e_{1,4k}, \ldots, e_{t,4k}, \ldots, e_{k+1,4k}, \ldots, e_{k+t,4k}, \ldots, e_{2k+1,4k}, \ldots, e_{2k+t,4k}, \ldots, e_{3k+1,4k}, \ldots, e_{3k+t,4k}\}.$$

Then for any $x \in \mathcal{X}_{t,k}$, $A_k x, \nabla h(A_k x) \in \mathcal{Y}_{t,k}$ and $A_k^T \nabla h(A_k x)$, $\nabla f_k(x) \in \mathcal{X}_{t+1,k}$. Furthermore,

$$\min_{x \in \mathcal{X}_{t,k}} f_k(x) = 8(k-t)\log 2 + \min_{u \in \mathbb{R}^t} f_t(u). \tag{4.9}$$

Proof. Let $x \in \mathcal{X}_{t,k}$. Thus, x can be decomposed into $x = (0_{k-t}^T, u^T)^T$ for $u \in \mathbb{R}^t$. Consequently,

$$A_k x = \begin{pmatrix} 2\sigma W_t u \\ 0_{k-t} \\ -2\zeta W_t u \\ 0_{k-t} \\ -2\sigma W_t u \\ 0_{k-t} \\ 2\zeta W_t u \\ 0_{k-t} \end{pmatrix},\ \nabla h(A_k x) = \begin{pmatrix} \tanh(\sigma W_t u) \\ 0_{k-t} \\ \tanh(-\zeta W_t u) \\ 0_{k-t} \\ \tanh(-\sigma W_t u) \\ 0_{k-t} \\ \tanh(\zeta W_t u) \\ 0_{k-t} \end{pmatrix}.$$

Hence, $A_k x, \nabla h(A_k x) \in \mathcal{Y}_{t,k}$. Noting that, since tanh is an odd function, we have

$$A_k^T \nabla h(A_k x) = 4\sigma W_k^T \begin{pmatrix} \tanh(\sigma W_t u) \\ 0_{k-t} \end{pmatrix} + 4\zeta W_k^T \begin{pmatrix} \tanh(\zeta W_t u) \\ 0_{k-t} \end{pmatrix}$$

and consequently

$$\nabla f_k(x) = A_k^T \nabla h(A_k x) - A_k^T b_k \in \mathcal{X}_{t+1,k}.$$

To show (4.9) note that, for $x \in \mathcal{X}_{t,k}$,

$$h(A_k x) = \sum_{i=1}^{t} 2\log(2\cosh(\sigma W_t u)_{(i)}) + 2\log(2\cosh(-\zeta W_t u)_{(i)}) \\ + 2\log(2\cosh(-\sigma W_t u)_{(i)}) + 2\log(2\cosh(\zeta W_t u)_{(i)}) + 8(k-t)\log\cosh(0)).$$

Applying again the definition of W_k in (1.3) here, we have $h(A_k x) = 8(k-t)\log 2 + h(A_t u)$. Moreover, since

$$b_k^T A_k x = 4(\sigma - \zeta)u = b_t^T A_t u$$

from the definitions in (3.8), we conclude the result (4.9) immediately. □

Similar to the discussion around Proposition 3.1, as a consequence of the previous lemma, we show that a first order method $\mathcal{M}$ satisfying the linear span assumption generates iterates $x_t \in \mathcal{X}_{t,k}$ when minimizing $f_k(x)$.

Lemma 4.4. Suppose that $\mathcal{M}$ is a deterministic first order method whose iterates satisfy the linear span assumption (4.2). When $\mathcal{M}$ is applied to minimize f_k defined in (4.4), we have $x_t \in \mathcal{X}_{t,k}$ for $1 \leq t \ \leq k$.

Proof. We will proceed via induction. Consider the case $t = 1$. By (4.2), $x_1 \in \text{span}\{\nabla f_k(x_0)\}$. Since $x_0 = 0$, $\nabla f_k(x_0) = -A_k^T b_k$. Thus, in view of (4.8),

$$\nabla f_k(x_0) \in \text{span}\{e_{k,k}\} = \mathcal{X}_{1,k}$$

so the statement holds for $t = 1$. Continuing with induction, assume that $x_i \in \mathcal{X}_{i,k}$ for $1 \leq i \leq t < k$. Noting lemma 4.2, $\nabla f_k(x_i) \in \mathcal{X}_{i+1,k}$ for all t. Thus, by (4.2), it follows that

$$x_{s+1} \in \text{span}\{\nabla f_k(x_0), \dots, \nabla f_k(x_s)\} \subseteq \mathcal{X}_{t+1,k}$$

which completes the induction process. □

Combining Lemmas 4.3, 4.4, and 4.2, it follows that

$$f_k(x_t) - f_k^* \geq \min_{x \in \mathcal{X}_{t,k}} f_k(x) - f_k^* \tag{4.10}$$

$$= 8(k-t)\log 2 + f_t^* - f_k^* \tag{4.11}$$

$$= 4(k-t)\left((\sigma - \zeta)c - \log\cosh(\sigma c) - \log\cosh(\zeta c)\right). \tag{4.12}$$

In an attempt to simplify the lower complexity bound presented above, we present the following lemma.

Lemma 4.5. For any real numbers σ and ζ that satisfy $\sigma/\zeta = 1.3$,

$$(\sigma - \zeta)c - \log\cosh(\sigma c) - \log\cosh(\zeta c) \geq \frac{c^2\sigma^2}{2}.$$

Proof. For $c > 0$, the function $c \to c\tanh(c)$ is increasing since its derivative is positive. Thus, since $\zeta c < \sigma c$,

$$\zeta c \tanh(\zeta c) \leq \sigma c \tanh(\sigma c)$$

and consequently

$$c \tanh(\zeta c) \leq c \tanh(\sigma c).$$

Applying these inequalities to bound (4.7) gives

$$2\zeta \tanh(\zeta c) \leq \sigma - \zeta \leq 2\sigma \tanh(\sigma c)$$

Furthermore, because function $c \mapsto \tanh(c)$ has positive derivative everywhere, we conclude that $c \in [a, b]$ where

$$a = \frac{1}{\sigma}\operatorname{arctanh}\left(\frac{1}{2} - \frac{\zeta}{2\sigma}\right),\ b = \frac{1}{\zeta}\operatorname{arctanh}\left(\frac{\sigma}{2\zeta} - \frac{1}{2}\right). \tag{4.13}$$

Here, since $\sigma/\zeta = 1.3$, we have $a, b > 0$. Thus, applying (4.7) we have

$$(\sigma - \zeta)c - \log\cosh(\sigma c) - \log\cosh(\zeta c) = \frac{(c\sigma)^2}{(c\sigma)^2}\left(\sigma c \tanh(\sigma c) + \zeta c \tanh(\zeta c) - \log\cosh(\sigma c) - \log\cosh(\zeta c)\right).$$

Noting again by the first derivative that function $c \mapsto c\tanh(c) - \log\cosh(c)$ is increasing for $c > 0$, we may apply (4.13) and obtain

$$(\sigma - \zeta)c - \log\cosh(\sigma c) - \log\cosh(\zeta c) \geq c^2\sigma^2 C \tag{4.14}$$

where

$$C := \frac{1}{(b\sigma)^2}\left(\sigma a \tanh(\sigma a) - \log\cosh(\sigma a) + \zeta a \tanh(\zeta a) - \log\cosh(\zeta a)\right).$$

Since C only depends on $\sigma/\zeta = 1.3$, we may numerically verify that

$$C > \frac{1}{2} \tag{4.15}$$

and the statement of the lemma follows from (4.14) and (4.15) immediately. □

We are now ready to estimate the lower complexity bound of first order methods applied to binary logistic regression under the linear span assumption.

Theorem 4.1. Let $\mathcal{M}$ be a deterministic first order method applied to solve (BLR) whose iterates satisfy the linear span assumption (4.2). For any iteration count M and constants $n = 2M$, $N = 8M$, there exist data matrix $A \in \mathbb{R}^{N\times n}$ and response vector $b \in \{-1, 1\}^N$ such that the M-th iterate generated by $\mathcal{M}$ satisfies

$$\phi_{A,b}(x_M) - \phi_{A,b}(x^*) \geq \frac{3\,||A||^2\,||x_0 - x^*||^2}{24(2M+1)(4M+1)} \tag{4.16}$$

and

$$||x_M - x^*||^2 > \frac{1}{8}\,||x_0 - x^*||^2. \tag{4.17}$$

Proof. We begin by setting constants $\zeta > 0$ and $\sigma = 1.3\zeta$ and defining A_k as in (3.8). Let $\mathcal{M}$ be applied to minimize f_k defined in (4.4) with $k = 2M$. Then from Lemma 4.4, the discussion in after (4.10), and the simplification in Lemma 4.5, we have

$$f_k(x_t) - f_k^* \geq 2(k-t)c^2\sigma^2,\ \forall t \leq k \tag{4.18}$$

Furthermore, from (4.6) and Lemma 4.2 along with the assumption that $x_0 = 0$, we have

$$||A_k|| \leq 4\sqrt{2\sigma^2 + 2(\sigma/1.3)^2} < 8\sigma \tag{4.19}$$

and

$$||x_0 - x^*||^2 = c^2 \sum_{i=1}^{k} i^2 = \frac{c^2}{6} k(k+1)(2k+1). \tag{4.20}$$

Combining the above two relations with (4.10), we conclude

$$f_k(x_t) - f_k^* > \frac{3(k-t)\,||A_k||^2\,||x_0 - x^*||^2}{16k(k+1)(2k+1)}. \tag{4.21}$$

Continuing, since $x_t \in \mathcal{X}_{t,k}$ by Lemma 4.4, we have the bound

$$||x_t - x^*||^2 \geq c^2 \sum_{i=1}^{k-t} i^2 = \frac{c^2}{6}(k-t)(k-t+1)(2k-2t+1). \tag{4.22}$$

Consequently, by setting $t = M$ and noting $k = 2M$, (4.21) becomes

$$f_k(x_M) - f_k^* > \frac{3\,||A_k||^2\,||x_0 - x^*||^2}{24(2M+1)(4M+1)}.$$

Applying (4.20) and (4.22) results in

$$||x_M - x^*||^2 \leq \frac{c^2}{6}M(M+1)(2M+1) > \frac{c^2}{48}(2M)(2M+1)(4M+1) = \frac{1}{8}\,||x_0 - x^*||^2\,.$$

Letting $A := A_k$ and $b := b_k$, we conclude the theorem. □

4.2 Lower Complexity Bound for solving Binary Logistic Regression via General first order Methods

In this section we will remove the assumption in (4.2) and build on the results of Theorem 4.1 to establish a similar result without the linear span assumption.

Lemma 4.6. For A_k, b_k specified in (3.8), any first order method $\mathcal{M}$, and some $t \leq \frac{k-3}{2}$, there exists an orthogonal matrix $U_t \in \mathbb{R}^{k \times k}$ satisfying

1. $U_t A_k^T b_k = A_k^T b_k$
2. When $\mathcal{M}$ is applied to solve (HBLR), the iterates $x_0, \ldots, x_t$ satisfy

$$x_i \in U_t^T \mathcal{X}_{2i+1,k},\ i = 0, \ldots, t.$$

Proof. Let us first define the set

$$\mathcal{U} := \{V \in \mathbb{R}^{k\times k} V \mid \text{ is orthogonal and } V A_k^T b_k = A_k^T b_k\}.$$

We will proceed by induction. The case $t = 0$ follows immediately from letting U_0 be the identity matrix. Now suppose the statement holds with $t = s-1 < (k-1)/2$ and denote x_s the next iterate. We will show the statement holds with $t = s$. Since $s < (k-1)/2$ is sufficiently small, it follows that $\mathcal{X}_{2s,k} \subset \mathcal{X}_{2s+1,k}$ and consequently $U_{s-1}^T \mathcal{X}_{2s,k} \subset U_{s-1}^T \mathcal{X}_{2s+1,k}$. Moreover, by Lemma 3.2, there exists some orthogonal matrix V such that

$$Vx = x \ \forall x \in U_{s-1}^T \mathcal{X}_{2s,k} \text{ and } Vx_s \in U_{s-1}^T \mathcal{X}_{2s+1,k}. \tag{4.23}$$

Denote then

$$U_s = U_{s-1}V. \tag{4.24}$$

From computation in (4.8), we know that $A_k^T b_k \in \mathcal{X}_{1,k} \subset \mathcal{X}_{2s,k}$. Thus, we have $U_s^T A_k^T b_k = V^T U_{s-1}^T A_k^T b_k = V^T A_k^T b_k = A_k^T b_k$. Noting that V is the product of orthogonal matrices and thus orthogonal, we conclude that $U_s \in \mathcal{U}$. Now suppose $x \in U_s^T \mathcal{X}_{2s-1,k}$. By (4.23) and the fact that $U_s \in \mathcal{U}$, we have

$$l_{A_k U_s, b_k}(x) = h(A_k U_s) - x^T U_s^T A_k^T b_k = h(A_k U_{s-1} x) - x^T U_{s-1}^T A_k^T b_k = l_{A_k U_{s-1}, b_k}(x).$$

Additionally, applying Lemma 4.2 and (4.23), we see that $V^T U_{s-1}^T A_k^T \nabla h(A_k U_{s-1} x) = U_{s-1}^T A_k^T \nabla h(A_k U_{s-1} x)$ and therefore,

$$\nabla l_{A_k U_s, b_k}(x) = U_s^T A_k^T \nabla h(A_k U_s x) - U_s^T A_k^T b_k = U_{s-1}^T A_k^T \nabla h(A_k U_s x) - U_{s-1}^T A_k^T b_k = \nabla l_{A_k U_{s-1}, b_k}(x).$$

Therefore, for any $x \in U_s^T \mathcal{X}_{2s-1,k}$, the first order method $\mathcal{M}$ receives same information regardless whether it is applied to $l_{A_k U_s, b_k}$ or $l_{A_k U_{s-1}, b_k}$. Thus, it produces the same iterates $x_0, \ldots, x_{s-1}$ when minimizing $l_{A_k U_s, b_k}$ as it does when minimizing $l_{A_k U_{s-1}, b_k}$. Furthermore, by the construction of V, for any $i = 0, \ldots, s-1$, we have

$$U_s^T \mathcal{X}_{2i+1,k} = V^T U_{s-1}^T \mathcal{X}_{2i+1,k} = U_{s-1}^T \mathcal{X}_{2i+1,k} \tag{4.25}$$

and

$$x_s \in V^T U_{s-1}^T \mathcal{X}_{2i+1,k} = U_s^T \mathcal{X}_{2s+1,k}.$$

Combining these two arguments with the induction hypothesis yields

$$x_i \in U_s^T \mathcal{X}_{2i+1,k},\ \forall i = 0, \dots, s. \tag{4.26}$$

In view of (4.26), we conclude the induction for the $t = s$ case by choosing $U = U_s$. □

Theorem 4.2. For any first order method $\mathcal{M}$ and fixed iteration number M with corresponding constants $N = 10M+8, n = 4M+2$, there always exists data matrix $A \in \mathbb{R}^{N\times n}$ and response vector $b \in \mathbb{R}^N$ such that when $\mathcal{M}$ is applied to solve (HBLR), the M-th iterate satisfies

$$l_{A,b}(x_M) - l_{A,b}^* \geq \frac{3\,||A||^2\,||x_0 - x^*||^2}{16(4M+3)(8M+5)}$$

and

$$||x_M - z^*||^2 > \frac{1}{8}\,||x_0 - z^*||^2$$

where z^* is the minimizer of $l_{A,b}$.

Proof. Let $\zeta > 0$ and set $\sigma = 1.3\zeta$. Let $k = 4M+2$ and define A_k using σ, ζ as in (3.8). Lemma 4.6 provides an orthogonal matrix U such that $U^T A_k^T b_k = A_k^T b_k$ and when $\mathcal{M}$ is applied to solve l_{A_kU,b_k}, the iterates x_i satsify $x_i \in U^T \mathcal{X}_{2i+1,k}$ for any $0 \leq i \leq M$. Thus, we have

$$\begin{aligned}
l_{A_kU,b_k}(x_M) &\geq \min_{x\in U^T\mathcal{X}_{2M+1,k}} l_{A_kU,b_k}(x)\\
&= \min_{x\in U^T\mathcal{X}_{2M+1,k}} h(A_kUx) - x^T U^T A_k^T b_k\\
&= \min_{x\in \mathcal{X}_{2M+1,k}} h(A_kx) - x^T A_k^T b_k\\
&= \min_{x\in \mathcal{X}_{2M+1,k}} f_k(x)
\end{aligned}$$

and

$$\begin{aligned} l^*_{A_kU,b_k} &= \min_{x\in\mathbb{R}^k} l_{A_kU,b_k}(x) \\ &= \min_{x\in\mathbb{R}^k} h(A_kUx) - x^TU^TA_k^Tb_k \\ &= \min_{x\in\mathbb{R}^k} h(A_kx) - x^TA_k^Tb_k \\ &= \min_{x\in\mathbb{R}^k} f_k(x). \end{aligned}$$

Here, we used the definition of f_k as in (4.4). Noting the two above computations applying (4.10) and Lemma 4.5, we see that

$$\begin{aligned} l_{A_kU,b_k}(x_T) - l^*_{A_kU,b_k} &\geq \min_{x\in\mathcal{X}_{2M+1,k}} f_k(x) - \min_{x\in\mathcal{X}_{2M+1,k}} f_k(x) \\ &= 4(k-2M-1)((\sigma-\zeta)c - \log\cosh(\sigma c) - \log\cosh(\zeta c)) \\ &\geq 2(k-2M-1)c^2\sigma^2. \end{aligned} \tag{4.27}$$

In view of computation of $l^*_{A_kU,b_k}$, we see that its minimizer z^* must satisfy $z^T = U^Tx^*$ with x^* being the minimizer of f_k provided in Lemma 4.2. Since Lemma 4.6 guarantees $x_M \in U^T\mathcal{X}_{2M+1,k}$, it follows that

$$\begin{aligned} ||x_M - z^*|| &\geq \max_{x\in\mathcal{X}_{2M+1,k}} ||x-x^*||^2 \\ &\geq c^2 \sum_{i=1}^{k-2M-1} i^2 \\ &\geq c^2 \sum_{i=1}^{2M+1} i^2 \\ &= \frac{c^2}{6}(2M+1)(2M+2)(4M+3) \end{aligned}$$

since we set $k = 4M+2$. Furthermore, because $x_0 = 0$,

$$||x_0 - z^*||^2 = \left|\left|U^Tx^*\right|\right|^2 = ||x^*||^2 = c^2\sum_{i=1}^{4M+2} i^2 = \frac{c^2}{6}(4M+2)(4M+3)(8M+5) \tag{4.28}$$

and consequently

$$||x_M - z^*||^2 > \frac{1}{8}\,||x_0 - z^*||\,. \tag{4.29}$$

Lastly, applying (4.28) to (4.27) and noting $k = 4M + 2$, we see that

$$l_{A_k U, b_k}(x_M) - l^*_{A_k U, b_k} \geq \frac{6\sigma^2(2M+1)\,||x_0 - z^*||^2}{(2M+1)(4M+3)(8M+5)}.$$

Recalling (4.6) we conclude that

$$l_{A_k U, b_k}(x_M) - l^*_{A_k U, b_k} \geq \frac{6\sigma^2\,||A||^2\,||x_0 - z^*||^2}{||A||^2\,(4M+3)(8M+5)} \geq \frac{6\sigma^2\,||A||^2\,||x_0 - z^*||^2}{32\sigma^2(4M+3)(8M+5)} = \frac{3\,||A||^2\,||x_0 - z^*||^2}{16(4M+3)(8M+5)}.$$

along with (4.29) completes the proof by setting $A := A_k U$ and $b := b_k$. □

4.3 Conclusions

With Theorem 4.2, we are able to construct a homogeneous worst case dataset for binary logistic regression which proves that Algorithm 1 is an optimal first order method for such a class of functions. However, the dataset generated here is most definitely non-standard. Possible extensions of this work would be to substitute our contrived A_k, b_k in favor of randomly generated datasets for more practical results.

Bibliography

[1] Y. E. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course.* Kluwer Academic Publishers, Massachusetts, 2004.

[2] Yuyuan Ouyang and Yangyang Xu. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. *Mathematical Programming*, pages 1–35, 2019.

[3] L. Pasquini S. Noschese and L. Reichel. Tridiagonal toeplitz matrices: properties and novel applications. *Numerical Linear Algebra with Applications*, 2012.

[4] A. Nemirovski and D. Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience Series in Discrete Mathematics. John Wiley, XV, 1983.

[5] Blake Woodworth, Jialei Wang, Brendan McMahan, and Nathan Srebro. Graph oracle models, lower bounds, and gaps for parallel stochastic optimization. *arXiv preprint arXiv:1805.10222*, 2018.

[6] A. S. Nemirovski. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.