New Results for Sparsity-inducing Methods for Logistic Regression

Robert M. Freund (MIT)

joint with Paul Grigas (Berkeley) and Rahul Mazumder (MIT)

SIOPT Vancouver, May 2017
How can optimization inform statistics (and machine learning)?

Paper in preparation (this talk):

*New Results for Sparsity-inducing Methods for Logistic Regression*

A “cousin” paper of ours:

*A New Perspective on Boosting in Linear Regression via Subgradient Optimization and Relatives*
Outline

- Optimization primer: some “old” results and new observations for Greedy Coordinate Descent (GCD)

- Logistic regression perspectives: statistics and machine learning

  - When the sample data is non-separable:
    - a “condition number” for the degree of non-separability
    - informing the convergence properties of GCD
    - reaching linear convergence of GCD (thanks to Bach)

  - When the sample data is separable:
    - a “condition number” for the degree of separability of the data
    - informing convergence to a certificate of separability

- Under construction: a different convergence result for an “accelerated” (but non-sparse) method for logistic regression (thanks to Renegar)
Some “Old” Results and New Observations for the Greedy Coordinate Descent Method
Greedy Coordinate Descent

\[ F^* := \min_x F(x) \quad \text{s.t.} \quad x \in \mathbb{R}^p \]

Greedy Coordinate Descent

Initialize at \( x^0 \in \mathbb{R}^p, k \leftarrow 0 \)

At iteration \( k \) :

1. Compute gradient \( \nabla F(x^k) \)
2. Compute
   - \( j_k \in \arg \max_{j \in \{1, \ldots, p\}} \{ |\nabla F(x^k)_j| \} \) and
   - \( d^k \leftarrow \text{sgn}(\nabla F(x^k)_{j_k})e_{j_k} \)
3. Choose step-size \( \alpha_k \)
4. Set \( x^{k+1} \leftarrow x^k - \alpha_k d^k \)
Greedy Coordinate Descent $\equiv \ell_1$-Steepest Descent

$$F^* := \min_x F(x) \quad \text{s.t. } x \in \mathbb{R}^p$$

Steepest Descent method in the $\ell_1$-norm

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

At iteration $k$ :

1. Compute gradient $\nabla F(x^k)$
2. Compute direction: $d^k \leftarrow \arg \max_d \{\nabla F(x^k)^T d : \|d\|_1 \leq 1\}$
3. Choose step-size $\alpha_k$
4. Set $x^{k+1} \leftarrow x^k - \alpha_k d^k$
Greedy Coordinate Descent \( \equiv \ell_1\)-Steepest Descent, cont.

\[
d^k \in \arg \max_{\|d\|_1 \leq 1} \{ \nabla F(x^k)^T d \}
\]
**Computational Guarantees for Greedy Coordinate Descent**

\[
F^* := \min_{x} F(x) \quad \text{s.t.} \quad x \in \mathbb{R}^p
\]

Assume \(F(\cdot)\) is convex and \(\nabla F(\cdot)\) is Lipschitz with parameter \(L_F:\)

\[
\|\nabla F(x) - \nabla F(y)\|_\infty \leq L_F \|x - y\|_1 \quad \text{for all } x, y \in \mathbb{R}^p
\]

Two sets of interest:

\(S_0 := \{x \in \mathbb{R}^p : F(x) \leq F(x^0)\}\) is the level set of the initial point \(x^0\)

\(S^* := \{x \in \mathbb{R}^p : F(x) = F^*\}\) is the set of optimal solutions
Metrics for Evaluating Greedy Coordinate Descent, cont.

\[ S_0 := \{ x \in \mathbb{R}^p : F(x) \leq F(x^0) \} \text{ is the level set of the initial point } x^0 \]

\[ S^* := \{ x \in \mathbb{R}^p : F(x) = F^* \} \text{ is the set of optimal solutions} \]

\[
\text{Dist}_0 := \max_{x \in S_0} \min_{x^* \in S^*} \| x - x^* \|_1
\]

(In high-dimensional machine learning problems, \( S^* \) can be very big)
Computational Guarantees for Greedy Coordinate Descent

\[ \text{Dist}_0 := \max_{x \in S_0} \min_{x^* \in S^*} \| x - x^* \|_1 \]

**Theorem: Objective Function Value Convergence (essentially [Beck and Tetruashvil 2014])**

If the step-sizes are chosen using the rule:

\[ \alpha_k = \frac{\| \nabla F(x^k) \|_\infty}{L_F} \]

for all \( k \geq 0 \),

then for each \( k \geq 0 \) the following inequality holds:

\[
F(x^k) - F^* \leq \frac{1}{F(x^0) - F^*} + \frac{k}{2L_F(\text{Dist}_0)^2} < \frac{2L_F(\text{Dist}_0)^2}{k}.
\]

Note that \( \alpha_k \to 0 \) as \( \| \nabla F(x^k) \|_\infty \to 0 \)
Theorem: Gradient Norm Convergence

For any step-size sequence \( \{\alpha_k\} \) and for each \( k \geq 0 \), it holds that:

\[
\min_{i \in \{0, \ldots, k\}} \| \nabla F(x^i) \|_\infty \leq \frac{F(x^0) - F^* + \frac{L_F}{2} \sum_{i=0}^{k} \alpha_i^2}{\sum_{i=0}^{k} \alpha_i}.
\]

If the step-sizes are chosen using the rule:

\[
\alpha_k = \frac{\| \nabla F(x^k) \|_\infty}{L_F} \quad \text{for all } k \geq 0,
\]

then for each \( k \geq 0 \) the following inequality holds:

\[
\min_{i \in \{0, \ldots, k\}} \| \nabla F(x^i) \|_\infty \leq \sqrt{\frac{2L_F(F(x^0) - F^*)}{k + 1}}.
\]
Theorem: Iterate Shrinkage

For any step-size sequence \( \{\alpha_k\} \), it holds for each \( k \geq 0 \) that:

\[
\|x^k\|_1 \leq \|x^0\|_1 + \sum_{i=0}^{k-1} \alpha_i.
\]

If the step-sizes are chosen using the rule:

\[
\alpha_k = \frac{\|\nabla F(x^k)\|_\infty}{L_F} \quad \text{for all } k \geq 0,
\]

then for each \( k \geq 0 \) it holds that:

\[
\|x^k\|_1 \leq \|x^0\|_1 + \sqrt{k} \sqrt{\frac{2(F(x^0) - F^*)}{L_F}}.
\]
Logistic Regression

- statistics perspective
- machine learning perspective
Logistic Regression Statistics Perspective
Example: Predicting Parole Violation

Predict $P$(violate parole) based on age, gender, time served, offense class, multiple convictions, NYC, etc.

<table>
<thead>
<tr>
<th>Violator</th>
<th>Male</th>
<th>Age</th>
<th>TimeServed</th>
<th>Class</th>
<th>Multiple</th>
<th>InCity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>49.4</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>26.0</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>24.9</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>1</td>
<td>52.1</td>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>1</td>
<td>35.9</td>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>25.9</td>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>1</td>
<td>19.0</td>
<td>D</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>1</td>
<td>43.2</td>
<td>C</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>1</td>
<td>31.6</td>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>1</td>
<td>40.7</td>
<td>B</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>1</td>
<td>53.9</td>
<td>A</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>1</td>
<td>28.5</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>1</td>
<td>36.1</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>1</td>
<td>48.8</td>
<td>D</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>1</td>
<td>37.6</td>
<td>C</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>1</td>
<td>42.5</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6098</td>
<td>0</td>
<td>1</td>
<td>55.0</td>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6099</td>
<td>0</td>
<td>1</td>
<td>49.6</td>
<td>A</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6100</td>
<td>0</td>
<td>1</td>
<td>22.4</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6101</td>
<td>0</td>
<td>1</td>
<td>44.8</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6102</td>
<td>0</td>
<td>0</td>
<td>45.3</td>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Logistic Regression for Prediction

$Y \in \{-1, 1\}$ is a Bernoulli random variable:

\[
P(Y = 1) = p
\]

\[
P(Y = -1) = 1 - p
\]

$x = (x_1, \ldots, x_p) \in \mathbb{R}^p$ is the vector of independent variables

$P(Y = 1)$ depends on the values of the independent variables $x_1, \ldots, x_p$

Logistic regression model is:

\[
P(Y = 1 \mid x) = \frac{1}{1 + e^{-\beta^T x}}
\]
Logistic Regression for Prediction, continued

Logistic regression model is:

\[ P(Y = 1 \mid x) = \frac{1}{1 + e^{-\beta^T x}} \]

Data records are \((x_i, y_i), i = 1, \ldots, n\)

<table>
<thead>
<tr>
<th>Violator</th>
<th>Male</th>
<th>Age</th>
<th>TimeServed</th>
<th>Class</th>
<th>Multiple</th>
<th>InCity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>149.4</td>
<td>3.15</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>126.0</td>
<td>5.95</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>124.9</td>
<td>2.25</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>152.1</td>
<td>29.22</td>
<td>A</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>135.9</td>
<td>12.78</td>
<td>A</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>125.9</td>
<td>1.18</td>
<td>C</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
<td>119.0</td>
<td>0.54</td>
<td>D</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>143.2</td>
<td>1.07</td>
<td>C</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>131.6</td>
<td>1.17</td>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>140.7</td>
<td>4.64</td>
<td>B</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>153.9</td>
<td>21.61</td>
<td>A</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>128.5</td>
<td>3.23</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>136.1</td>
<td>3.71</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>148.8</td>
<td>1.17</td>
<td>D</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>0</td>
<td>137.6</td>
<td>4.62</td>
<td>C</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>0</td>
<td>142.5</td>
<td>1.75</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6098</td>
<td>0</td>
<td>155.0</td>
<td>0.72</td>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6099</td>
<td>0</td>
<td>149.6</td>
<td>29.88</td>
<td>A</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6100</td>
<td>0</td>
<td>122.4</td>
<td>2.85</td>
<td>D</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6101</td>
<td>0</td>
<td>144.8</td>
<td>1.76</td>
<td>D</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6102</td>
<td>0</td>
<td>45.3</td>
<td>1.03</td>
<td>E</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Let us construct an estimate of \(\beta\) based on the data \((x_i, y_i), i = 1, \ldots, n\)
Logistic Regression: Maximum Likelihood Estimation

\[
\begin{align*}
\max_\beta \left( \prod_{y_i=1} \frac{1}{1 + e^{-\beta^T x_i}} \right) \left( \prod_{y_i=-1} \left(1 - \frac{1}{1 + e^{-\beta^T x_i}}\right) \right) \\
= \max_\beta \left( \prod_{y_i=1} \frac{1}{1 + e^{-\beta^T x_i}} \right) \left( \prod_{y_i=-1} \frac{1}{1 + e^{\beta^T x_i}} \right) \\
= \max_\beta \left( \prod_{i=1}^n \frac{1}{1 + e^{-y_i \beta^T x_i}} \right) \\
\equiv \min_\beta \frac{1}{n} \sum_{i=1}^n \ln \left( 1 + e^{-y_i \beta^T x_i} \right) \quad =: \ L_n(\beta)
\end{align*}
\]
Logistic Regression: Maximum Likelihood Optimization Problem

Logistic regression optimization problem is:

\[ L^*_n := \min_{\beta} L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i \beta^T x_i)) \]
\[ \text{s.t. } \beta \in \mathbb{R}^p \]

The logistic term is a 1-smoothing of \( f(\alpha) = \max\{0, -\alpha\} \)
(\( \equiv \) shifted “hinge loss”)
Logistic Regression: Machine Learning Perspective
Logistic Regression as Binary Classification

Data: \((x_i, y_i) \in \mathbb{R}^p \times \{-1, 1\}, \ i = 1, \ldots, n\)
- \(x = (x_1, \ldots, x_p) \in \mathbb{R}^p\) is the vector of features (ind. variables)
- \(y \in \{-1, 1\}\) is the response/label

Task: predict \(y\) based on the linear function \(\beta^T x\)
- \(\beta \in \mathbb{R}^p\) are the model coefficients

Loss function: \(\ell(y, \beta^T x)\) represents the loss incurred when the truth is \(y\) but our classification/prediction was based on \(\beta^T x\)

**Loss Minimization Problem:** \(\min_\beta \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \beta^T x_i)\)
Loss Functions for Binary Classification

Some common loss functions used for binary classification

- **0-1 loss**: \( \ell(y, \beta^T x) := 1(y \beta^T x < 0) \)
- **Hinge loss**: \( \ell(y, \beta^T x) := \max(0, 1 - y \beta^T x) \)
- **Logistic loss**: \( \ell(y, \beta^T x) := \ln(1 + \exp(-y \beta^T x)) \)

Here “Margin” = \( y\beta^T x \)
Advantages of Logistic Loss Function

Why use the logistic loss function for classification?

- Computational advantages: convex, smooth
- Fits previous statistical model of conditional probability:
  \[ P(Y = y \mid x) = \frac{1}{1 + \exp(-y\beta^T x)} \]
- Makes sense when the data is **non-separable**
- Robust to misspecification of class labels
Alternate version of optimization problem adds regularization and/or sparsification:

\[
L^*_n := \min_{\beta} L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i \beta^T x_i)) + \lambda \|\beta\|_p
\]

\[
\text{s.t. } \beta \in \mathbb{R}^p
\]

\[
\|\beta\|_0 \leq k
\]

Aspirations:

- Good predictive performance on new (out of sample) observations
- Models that are more interpretable (e.g., sparse)
<table>
<thead>
<tr>
<th>GCD Primer</th>
<th>Logistic Regression</th>
<th>GCD for LR</th>
<th>Non-Separable Case</th>
<th>Separable Case</th>
<th>Other Issues</th>
</tr>
</thead>
</table>

**Greedy Coordinate Descent for Logistic Regression**
Greedy Coordinate Descent for Logistic Regression

Initialize at $\beta^0 \leftarrow 0, k \leftarrow 0$

At iteration $k \geq 0$:

1. Compute $\nabla L_n(\beta^k)$
2. Compute $j_k \in \arg\max_{j \in \{1, \ldots, p\}} |\nabla L_n(\beta^k)_j|$
3. Set $\beta^{k+1} \leftarrow \beta^k - \alpha_k \text{sgn}(\nabla L_n(\beta^k)_{j_k})e_{j_k}$

Why use Greedy Coordinate Descent for Logistic Regression?

- Scalable and effective when $n, p \gg 0$ and maybe $p > n$
- GCD performs variable selection
- GCD imparts implicit regularization
- Just one tuning parameter (number of iterations)
- Connections to boosting (LogitBoost)
Implicit Regularization and Variable Selection Properties

Artificial example: \( n = 1000, p = 100, \) true model has 5 non-zeros

Compare with explicit regularization schemes (\( \ell_1, \ell_2, \) etc.)
How Can GCD Inform Logistic Regression?

Some questions:

- How do the computational guarantees for Greedy Coordinate Descent specialize to the case of Logistic Regression?

- What role does problem structure/conditioning play in these guarantees?

- Can we say anything further about the convergence properties of Greedy Coordinate Descent in the special case of Logistic Regression?
**Basic Properties of the Logistic Loss Function**

$$L_n^* := \min_{\beta} L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i\beta^T x_i))$$

s.t. $\beta \in \mathbb{R}^p$

- $L_n(\cdot)$ is convex
- $\nabla L_n(\cdot)$ is $L = \frac{1}{4n} \|X\|_{1,2}^2$-Lipschitz:
  $$\|\nabla L_n(\beta) - \nabla L_n(\beta')\|_\infty \leq \frac{1}{4n} \|X\|_{1,2}^2 \|\beta - \beta'\|_1$$

  where $\|X\|_{1,2} := \max_{j=1,\ldots,p} \|X_j\|_2$

- For $\beta^0 := 0$ it holds that $L_n(\beta^0) = \ln(2)$
- $L_n^* \geq 0$
- If $L_n^* = 0$, then the optimum is not attained (something is “wrong” or “very wrong”)
- We will see later that “very wrong” is actually good....
Basic Properties, continued

$$L_n^* := \min_{\beta} \quad L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i \beta^T x_i))$$

Logistic regression “ideally” seeks $\beta$ for which $y_i x_i^T \beta > 0$ for all $i$:

- $y_i > 0 \Rightarrow x_i^T \beta > 0$
- $y_i < 0 \Rightarrow x_i^T \beta < 0$
Geometry of the Data: Separable and Non-Separable Data

(a) Separable Data
(b) Not Separable Data
(c) Mildly Non-Separable" Data
(d) Very Non-Separable Data
Separable Data

The data is separable if there exists $\bar{\beta}$ for which $y_i \cdot (\bar{\beta})^T x_i > 0$ for all $i = 1, \ldots, n$

Equivalently $Y \mathbf{x} \bar{\beta} > 0$ where $Y := \text{diag}(y)$
Separable Data, continued

Let
\[ L_n^* := \min_{\beta} L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i \beta^T x_i)) \]

The data is **separable** if there exists \( \bar{\beta} \) for which
\[ YX\bar{\beta} > 0 \quad \text{where} \quad Y := \text{diag}(y) \]

If \( \bar{\beta} \) separates the data, then \( L_n(\theta \bar{\beta}) \to 0 \) (\( = L_n^* \)) as \( \theta \to +\infty \)

Perhaps trying to optimize the logistic loss function is unlikely to be effective at finding a “good” linear separator?
Strictly Non-Separable Data

We say that the data is strictly non-separable if:

\[ YX\beta \neq 0 \Rightarrow YX\beta \not\geq 0 \]

(a) Strictly Non-Separable
(b) Not Strictly Non-Separable
Theorem: Attaining Optima

When the data is strictly non-separable, then the logistic regression problem attains its optimum.

Let us quantify the degree of non-separability of the data and relate this to problem behavior/conditioning.

(a) Mildly non-separable data
(b) Very non-separable data
Non-Separability Measure $\text{DistSEP}^*$

**Definition of Non-Separability Measure $\text{DistSEP}^*$**

$$\text{DistSEP}^* := \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} [y_i \beta^T x_i]^-$$

s.t. $\|\beta\|_1 = 1$

$\text{DistSEP}^*$ is the least average misclassification error

$\text{DistSEP}^* > 0$ if and only if the data is strictly non-separable
Non-Separability Measure $\text{DistSEP}^*$

$$\text{DistSEP}^* := \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} [y_i \beta^T x_i]^-$$

s.t. \quad \|\beta\|_1 = 1

(a) $\text{DistSEP}^*$ is small

(b) $\text{DistSEP}^*$ is large
DistSEP\(^*\) and “Distance to Separability”

\[
\text{DistSEP}^* := \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} [y_i \beta^T x_i]^- \\
\text{s.t.} \quad \|\beta\|_1 = 1
\]

Theorem: DistSEP\(^*\) is the “Distance to Separability”

\[
\text{DistSEP}^* = \inf_{\Delta x_1, \ldots, \Delta x_n} \frac{1}{n} \sum_{i=1}^{n} \|\Delta x_i\|_\infty \\
\text{s.t.} \quad (x_i + \Delta x_i, y_i), i = 1, \ldots, n \text{ are separable}
\]
DistSEP$^*$ and Problem Behavior/Conditioning

\[ L_n^* := \min_{\beta} L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i \beta^T x_i)) \]

\[ \text{DistSEP}^* := \min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} [y_i \beta^T x_i]^- \]

s.t. \( \|\beta\|_1 = 1 \)

**Theorem: Strict Non-Separability and Sizes of Optimal Solutions**

Suppose that the data is strictly non-separable, and let \( \beta^* \) be an optimal solution of the logistic regression problem. Then

\[ \|\beta^*\|_1 \leq \frac{\ln(2)}{\text{DistSEP}^*}, \quad \text{whereby} \quad \text{Dist}_0 \leq \frac{2 \ln(2)}{\text{DistSEP}^*}. \]
Consider Greedy Coordinate Descent applied to the Logistic Regression problem with step-sizes $\alpha_k := \frac{4n\|\nabla L_n(\beta^k)\|_\infty}{\|X\|_{1,2}^2}$ for all $k \geq 0$, and suppose that the data is strictly non-separable. Then for each $k \geq 0$ it holds that:

(i) (training error): $L_n(\beta^k) - L^*_n \leq \frac{2(\ln(2))^2\|X\|_{1,2}^2}{k \cdot n \cdot (\text{DistSEP}^*)^2}$

(ii) (gradient norm): $\min_{i \in \{0, \ldots, k\}} \|\nabla L_n(\beta^i)\|_\infty \leq \|X\|_{1,2} \sqrt{\frac{(\ln(2) - L^*_n)}{2n \cdot (k+1)}}$

(iii) (regularization): $\|\beta^k\|_1 \leq \sqrt{k} \left(\frac{1}{\|X\|_{1,2}}\right) \sqrt{8n(\ln(2) - L^*_n)}$

(iv) (sparsity): $\|\beta^k\|_0 \leq k$
Theorem: Computational Guarantees for GCD: Non-Separable Case

Consider Greedy Coordinate Descent applied to the Logistic Regression problem with step-sizes $\alpha_k := \frac{4n\|\nabla L_n(\beta^k)\|_{\infty}}{\|X\|_{1,2}^2}$ for all $k \geq 0$, and suppose that the data is strictly non-separable. Then for each $k \geq 0$ it holds that:

(i) (training error): $L_n(\beta^k) - L^*_n \leq \frac{2(\ln(2))^2\|X\|_{1,2}^2}{k \cdot n \cdot (\text{DistSEP}^*)^2}$

(ii) (gradient norm): $\min_{i \in \{0, \ldots, k\}} \|\nabla L_n(\beta^i)\|_{\infty} \leq \|X\|_{1,2} \sqrt{\frac{(\ln(2) - L^*_n)}{2n \cdot (k+1)}}$

(iii) (regularization): $\|\beta^k\|_1 \leq \sqrt{k} \left( \frac{1}{\|X\|_{1,2}} \right) \sqrt{8n(\ln(2) - L^*_n)}$

(iv) (sparsity): $\|\beta^k\|_0 \leq k$
Reaching Linear Convergence using Greedy Coordinate Descent for Logistic Regression

For logistic regression, does Greedy Coordinate Descent exhibit linear convergence?
Some Definitions/Notation

Definitions:

- \( R := \max_{i \in \{1, \ldots, n\}} \| x_i \|_2 \) (maximum norm of the feature vectors)
- \( H(\beta^*) \) denotes the Hessian of \( L_n(\cdot) \) at an optimal solution \( \beta^* \)
- \( \lambda_{\text{pmin}}(H(\beta^*)) \) denotes the smallest non-zero (and hence positive) eigenvalue of \( H(\beta^*) \)
Reaching Linear Convergence of GCD for Logistic Regression

**Theorem: Reaching Linear Convergence of GCD for Logistic Regression**

Consider Greedy Coordinate Descent applied to the Logistic Regression problem with step-sizes $\alpha_k := \frac{4n\|\nabla L_n(\beta^k)\|_{\infty}}{\|X\|_{1,2}^2}$ for all $k \geq 0$, and suppose that the data is strictly non-separable. Define:

$$\tilde{k} := \frac{16 \ln(2)^2 \|X\|_{1,2}^2 R^2 p}{9n(DistSE^{*})^2 \lambda_{\text{pmin}}(H(\beta^{*}))^2}.$$  

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

$$L_n(\beta^k) - L_n^* \leq (L_n(\beta^\tilde{k}) - L_n^*) \left(1 - \frac{\lambda_{\text{pmin}}(H(\beta^{*})n)}{\|X\|_{1,2}^2 p} \right)^{k-\tilde{k}}.$$
Some comments:

- Proof relies on (a slight generalization of) the “generalized self-concordance” property of the logistic loss function due to [Bach 2014]

- Furthermore, we can bound:

  $$\lambda_{p\min}(H(\beta^*)) \geq \frac{1}{4n} \lambda_{p\min}(X^TX) \exp\left(-\frac{\ln 2\|X\|1,\infty}{\text{DistSEP}^*}\right)$$

- As compared to results of a similar flavor for other algorithms, here we have an exact characterization of when the linear convergence “kicks in” and also what the rate of linear convergence is guaranteed to be

- Q: Can we exploit this generalized self-concordance property in other ways? (still ongoing . . .)
Separability and Problem Behavior/Conditioning

Separable data
Separable Data, continued

\[ L_n^* := \min_\beta \quad L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \ln(1 + \exp(-y_i \beta^T x_i)) \]

Recall the data is separable if there exists \( \bar{\beta} \) for which

\[ YX\bar{\beta} > 0 \quad \text{where} \quad Y := \text{diag}(y) \]

If \( \bar{\beta} \) separates the data, then \( L_n(\theta \bar{\beta}) \to 0 \ (= L_n^*) \) as \( \theta \to +\infty \)

Despite this, it turns out that GCD is reasonably effective at finding a “good” linear separator as we shall shortly see....
Margin function $\rho(\beta)$

$$\rho(\beta) := \min_{i \in \{1, \ldots, n\}} [y_i \beta^T x_i]$$
Separability Measure DistNSEP$^*$

Definition of Separability Measure DistNSEP$^*$

$$\text{DistNSEP}^* := \max_{\beta \in \mathbb{R}^p} \rho(\beta)$$

s.t. $\|\beta\|_1 = 1$

DistNSEP$^*$ is the maximum margin over all (normalized) $\beta$

DistNSEP$^*$ > 0 if and only if the data is separable
Separability Measure \( \text{DistNSEP}^* \)

\[
\text{DistNSEP}^* := \max_{\beta \in \mathbb{R}^p} \rho(\beta)
\]

s.t. \( \|\beta\|_1 = 1 \)
DistNSEP* and “Distance to Non-Separability”

\[
\text{DistNSEP}^* := \max_{\beta \in \mathbb{R}^p} \rho(\beta)
\]

s.t. \( \|\beta\|_1 = 1 \)

Theorem: DistNSEP* is the “Distance to Non-Separability”

\[
\text{DistNSEP}^* = \inf_{\Delta x_1, \ldots, \Delta x_n} \max_{i \in \{1, \ldots, n\}} \|\Delta x_i\|_\infty
\]

s.t. \((x_i + \Delta x_i, y_i), i = 1, \ldots, n\) are non-separable
Theorem: Computational Guarantees for GCD: Separable Case

Consider Greedy Coordinate Descent applied to the Logistic Regression problem with step-sizes \( \alpha_k := \frac{4n\|\nabla L_n(\beta^k)\|_{\infty}}{\|X\|_{1,2}^2} \) for all \( k \geq 0 \), and suppose that the data is separable.

(i) (margin bound): there exists \( i \leq \left\lfloor \frac{3.7n\|X\|_{1,2}^2}{(\text{DistNSEP}^*)^2} \right\rfloor \) for which the normalized iterate \( \bar{\beta}^i := \beta^i / \|\beta^i\|_1 \) satisfies

\[ \rho(\bar{\beta}^i) \geq \frac{.18 \cdot \text{DistNSEP}^*}{n}. \]

(ii) (gradient norm): \( \min_{i \in \{0, \ldots, k\}} \|\nabla L_n(\beta^i)\|_{\infty} \leq \|X\|_{1,2} \sqrt{\frac{(\ln(2) - L_n^*)}{2n(k+1)}} \)

(iii) (regularization): \( \|\beta^k\|_1 \leq \sqrt{k} \left( \frac{1}{\|X\|_{1,2}^2} \right) \sqrt{8n(\ln(2) - L_n^*)} \)

(iv) (sparsity): \( \|\beta^k\|_0 \leq k \)
Theorem: Computational Guarantees for GCD: Separable Case

Consider Greedy Coordinate Descent applied to the Logistic Regression problem with step-sizes \( \alpha_k := \frac{4n\|\nabla L_n(\beta_k)\|_\infty}{\|X\|_{1,2}^2} \) for all \( k \geq 0 \), and suppose that the data is separable.

(i) (margin bound): there exists \( i \leq \left\lfloor \frac{3.7n\|X\|_{1,2}^2}{(\text{DistNSEP}^*)^2} \right\rfloor \) for which the normalized iterate \( \bar{\beta}^i := \beta^i / \|\beta^i\|_1 \) satisfies
\[
\rho(\bar{\beta}^i) \geq \frac{0.18 \cdot \text{DistNSEP}^*}{n}.
\]

(ii) (gradient norm): \( \min_{i \in \{0, \ldots, k\}} \|\nabla L_n(\beta^i)\|_\infty \leq \|X\|_{1,2} \sqrt{\frac{(\ln(2) - L_n^*)}{2n(k+1)}} \)

(iii) (regularization): \( \|\beta_k\|_1 \leq \sqrt{k} \left( \frac{1}{\|X\|_{1,2}} \right) \sqrt{8n(\ln(2) - L_n^*)} \)

(iv) (sparsity): \( \|\beta^k\|_0 \leq k \)
Other Issues

Some other topics not mentioned today (still ongoing):

- Other “GCD-type”/“boosting-type” methods suggested by connections to Mirror Descent and the Frank-Wolfe method.

- High-dimensional regime $p > n$, define $\text{DistSEP}_k^*$ and $\text{DistNSEP}_k^*$ for restricting $\beta$ to satisfy $\|\beta\|_0 \leq k$.

- Numerical experiments comparing methods.

- Further investigation of the properties of other step-size choices for Greedy Coordinate Descent.
Summary

- Some “old” results and new observations for the Greedy Coordinate Descent Method

- Analyzing GCD for Logistic Regression: separable/non-separable cases

  - Non-Seperable case
    - behavioral/condition measure DistSEP*
    - computational guarantees for GCD including reaching linear convergence

  - Separable case
    - behavioral/condition measure DistNSEP*
    - computational guarantees for GCD including computing a reasonably good separator