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

Robert M. Freund (MIT)

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

ISI Marrakech, July 2017

How can optimization inform statistics (and machine learning)?

Paper in preparation (this talk):

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

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 the family of steepest descent algorithms

- Logistic regression perspectives: statistics and machine learning

- A pair of condition numbers for the logistic regression problem:
  - when the sample data is non-separable:
    - a condition number for the degree of non-separability of the dataset
    - informing the convergence guarantees of steepest descent family
    - guarantees on reaching linear convergence (thanks to Bach)
  - when the sample data is separable:
    - a condition number for the degree of separability of the dataset
    - informing convergence guarantee to deliver an approximate maximum margin classifier

Primer on Steepest Descent in a Given Norm

Some Old and New Results for Steepest Descent in a Given Norm

Steepest Descent in a Given Norm (SDGN)

$$\begin{array}{rccl} F^* & := & \min_x & F(x) \\ & & \text{s.t.} & x \in \mathbb{R}^p \end{array}$$

Let $\|\cdot\|$ be the given norm on the variables $x \in \mathbb{R}^p$

Steepest Descent in a Given Norm (SDGN)

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

At iteration k :

1. Compute gradient $\nabla F(x^k)$
2. Compute $d^k \leftarrow \arg\max_d \{\nabla F(x^k)^T d : \|d\| \leq 1\}$
3. Choose step-size α_k
4. Set $x^{k+1} \leftarrow x^k - \alpha_k d^k$

Greedy Coordinate Descent $\equiv \ell_1$-Steepest Descent

$$
\begin{array}{rlll}
F^* & := & \min_{x} & F(x) \\
 & & \text{s.t.} & x \in \mathbb{R}^p
\end{array}
$$

Let $\|\cdot\| = \|\cdot\|_1$

Steepest Descent method in the ℓ_1-norm

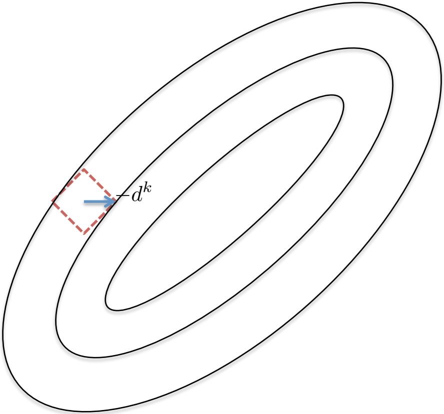
Initialize at $x^0 \in \mathbb{R}^p$, $k \leftarrow 0$

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 α_k
4. Set $x^{k+1} \leftarrow x^k - \alpha_k d^k$

Greedy Coordinate Descent $\equiv \ell_1$-Steepest Descent, cont.

$$d^k \in \arg \max_{\|d\|_1 \leq 1} \{\nabla F(x^k)^T d\}$$



Gradient Descent $\equiv \ell_2$-Steepest Descent

$$\begin{array}{rrcl} F^* & := & \min_x & F(x) \\ & & \text{s.t.} & x \in \mathbb{R}^p \end{array}$$

Let $\|\cdot\| = \|\cdot\|_2$

Steepest Descent method in the ℓ_2-norm

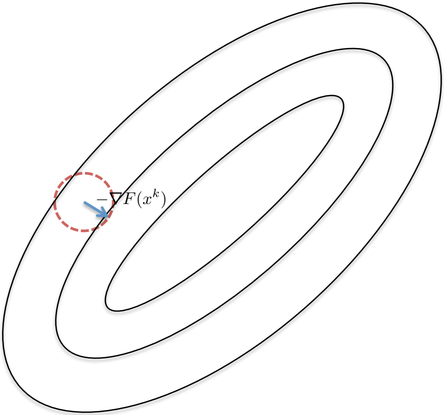
Initialize at $x^0 \in \mathbb{R}^p$, $k \leftarrow 0$

At iteration k :

1. Compute gradient $\nabla F(x^k)$
2. Compute direction: $d^k \leftarrow \arg\max_d \{\nabla F(x^k)^T d : \|d\|_2 \leq 1\}$
3. Choose step-size α_k
4. Set $x^{k+1} \leftarrow x^k - \alpha_k d^k$

Gradient Descent $\equiv$ ℓ_2-Steepest Descent, cont.

$$d^k \in \arg \max_{\|d\|_2 \leq 1} \{\nabla F(x^k)^T d\}$$



Computational Guarantees for Steepest Descent family

$$\begin{array}{rcll} F^* & := & \min\limits_x & F(x) \\ & & \text{s.t.} & x \in \mathbb{R}^p \end{array}$$

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

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

$\|\cdot\|_*$ is the usual dual norm

Two sets of interest:

$\mathcal{S}_0 := \{x \in \mathbb{R}^p : F(x) \leq F(x^0)\}$ is the level set of the initial point x^0

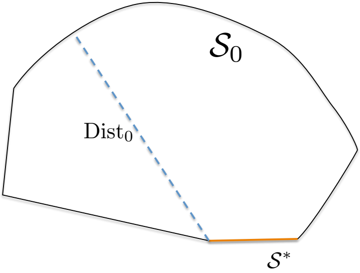
$\mathcal{S}^* := \{x \in \mathbb{R}^p : F(x) = F^*\}$ is the set of optimal solutions

Metrics for Evaluating Steepest Descent family, cont.

$\mathcal{S}_0 := \{x \in \mathbb{R}^p : F(x) \leq F(x^0)\}$ is the level set of the initial point x^0

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

$$\mathrm{Dist}_0 := \max_{x \in \mathcal{S}_0} \min_{x^* \in \mathcal{S}^*} \|x - x^*\|$$



(In high-dimensional machine learning problems, $\mathcal{S}^*$ can be very big)

Computational Guarantees for Steepest Descent family

$$\mathrm{Dist}_0 := \max_{x \in \mathcal{S}_0} \min_{x^* \in \mathcal{S}^*} \|x - x^*\|$$

Theorem: Objective Function Value Convergence (essentially [Beck and Tetruashvil 2014], [Nesterov 2003])

If the step-sizes are chosen using the rule:

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

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

$$F(x^k) - F^* \ \leq \ \frac{2L_F(\mathrm{Dist}_0)^2}{\hat{K}^0 + k} \ < \ \frac{2L_F(\mathrm{Dist}_0)^2}{k}$$

where $\hat{K}^0 := \dfrac{2L_F(\mathrm{Dist}_0)^2}{F(x^0) - F^*}$.

Computational Guarantees for Steepest Descent family, cont.

Theorem: Gradient Norm and Iterate Norm Convergence

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:

$$\|x^k - x^0\| \ \leq \ \sqrt{k}\sqrt{\frac{2(F(x^0) - F^*)}{L_F}} \ ,$$

and

$$\min_{i \in \{0,\ldots,k\}} \|\nabla F(x^i)\|_* \ \leq \ \sqrt{\frac{2L_F(F(x^0) - F^*)}{k+1}} \quad .$$

Logistic Regression

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.

```
           Violator Male  Age TimeServed Class Multiple InCity
1                 0    1 49.4       3.15     D        0      1
2                 1    1 26.0       5.95     D        1      0
3                 0    1 24.9       2.25     D        1      0
4                 0    1 52.1      29.22     A        0      0
5                 0    1 35.9      12.78     A        1      1
6                 0    1 25.9       1.18     C        1      1
7                 0    1 19.0       0.54     D        0      0
8                 0    1 43.2       1.07     C        0      1
9                 0    1 31.6       1.17     E        0      0
10                0    1 40.7       4.64     B        1      1
11                0    1 53.9      21.61     A        0      1
12                0    1 28.5       3.23     D        1      0
13                0    1 36.1       3.71     D        0      1
14                0    1 48.8       1.17     D        0      0
15                0    1 37.6       4.62     C        0      0
16                0    1 42.5       1.75     D        0      1
...             ...  ...  ...        ...   ...      ...    ...
6098              0    1 55.0       0.72     E        0      0
6099              0    1 49.6      29.88     A        0      1
6100              0    1 22.4       2.85     D        0      1
6101              0    1 44.8       1.76     D        1      0
6102              0    0 45.3       1.03     E        0      0
```

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:

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

Logistic Regression for Prediction, continued

Logistic regression model is:

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

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

```
     Violator Male  Age TimeServed Class Multiple InCity
1           0    1 49.4       3.15     D        0      1
2           1    1 26.0       5.95     D        1      0
3           0    1 24.9       2.25     D        1      0
4           0    1 52.1      29.22     A        0      0
5           0    1 35.9      12.78     A        1      1
6           0    1 25.9       1.18     C        1      1
7           0    1 19.0       0.54     D        0      0
8           0    1 43.2       1.07     C        0      1
9           0    1 31.6       1.17     E        0      0
10          0    1 40.7       4.64     B        1      1
11          0    1 53.9      21.61     A        0      1
12          0    1 28.5       3.23     D        1      0
13          0    1 36.1       3.71     D        0      1
14          0    1 48.8       1.17     D        0      0
15          0    1 37.6       4.62     C        0      0
16          0    1 42.5       1.75     D        0      1
...       ...  ...  ...        ...   ...      ...    ...
6098        0    1 55.0       0.72     E        0      0
6099        0    1 49.6      29.88     A        0      1
6100        0    1 22.4       2.85     D        0      1
6101        0    1 44.8       1.76     D        1      0
6102        0    0 45.3       1.03     E        0      0
```

Let us construct an estimate of β based on the data (x_i, y_i), $i = 1, \ldots, n$

Logistic Regression: Maximum Likelihood Estimation

$$\max_{\beta}\left(\prod_{y_i=1}\frac{1}{1+e^{-\beta^T x_i}}\right)\left(\prod_{y_i=-1}\left(1-\frac{1}{1+e^{-\beta^T x_i}}\right)\right)$$

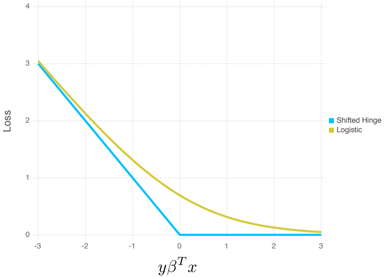
$$=\max_{\beta}\left(\prod_{i=1}^{n}\frac{1}{1+e^{-y_i\beta^T x_i}}\right)$$

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

Logistic Regression: Maximum Likelihood Optimization Problem

Logistic regression optimization problem is:

$$
\begin{array}{rccl}
L_n^* & := & \min_{\beta} & L_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ln(1+\exp(-y_i\beta^T x_i)) \\
 & & \text{s.t.} & \beta \in \mathbb{R}^p
\end{array}
$$



The logistic term is a 1-smoothing of $f(\alpha) = \max\{0, -\alpha\}$ ($\equiv$ shifted “hinge loss”)

Properties of the Logistic Loss Function

$$\begin{array}{rccl} L_n^* & := & \min_{\beta} & L_n(\beta) := \frac{1}{n}\sum_{i=1}^n \ln(1+\exp(-y_i\beta^T x_i)) \\ & & \text{s.t.} & \beta \in \mathbb{R}^p \end{array}$$

Proposition: Lipschitz constant of the gradient of $L_n(\beta)$

$\nabla L_n(\cdot)$ is $L = \frac{1}{4n}\|\mathbf{X}\|_{\cdot,2}^2$-Lipschitz:

$$\|\nabla L_n(\beta) - \nabla L_n(\beta')\|_* \leq \frac{1}{4n}\|\mathbf{X}\|_{\cdot,2}^2\|\beta - \beta'\|$$

where $\|\mathbf{X}\|_{\cdot,2} := \max_{\|\beta\|\leq 1} \|\mathbf{X}\beta\|_2$

Properties of the Logistic Loss Function, continued

$$
\begin{array}{rccl}
L_n^* & := & \min\limits_{\beta} & L_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-y_i\beta^T x_i)) \\
 & & \text{s.t.} & \beta \in \mathbb{R}^p
\end{array}
$$

- $L_n(\cdot)$ is convex
- $L_n^* \geq 0$
- 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 very good....

- For $\beta^0 := 0$ it holds that $L_n(\beta^0) = \ln(2)$

Logistic Regression: Machine Learning Perspective

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

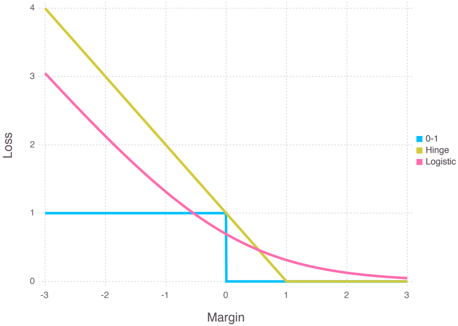
Loss function: $\ell(y, \beta^T x)$ represents the loss incurred when the truth is y but our classification/prediction was based on $\beta^T x$

Loss Minimization Problem: $\displaystyle \min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, \beta^T x_i)$

Loss Functions for Binary Classification

Some common loss functions used for binary classification

- 0-1 loss: $\ell(y, \beta^T x) := \mathbf{1}(y\beta^T x < 0)$
- 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 probablity:

$$P(Y = y \mid x) = \frac{1}{1+\exp(-y\beta^T x)}$$

- Makes sense when the data is non-separable
- Robust to misspecification of class labels

Logistic Regression Problem of Interest, continued

Alternate version of optimization problem adds regularization and/or sparsification:

$$\begin{array}{rccl} L_n^* & := & \min\limits_{\beta} & L_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-y_i\beta^T x_i)) + \lambda\|\beta\|_p \\ & & \text{s.t.} & \beta \in \mathbb{R}^p \\ & & & \\ & & & \|\beta\|_0 \le k \end{array}$$

Aspirations:

- Good predictive performance on new (out of sample) observations
- Models that are more interpretable (e.g., sparse)

Computational Experiment: Greedy Coordinate Descent (GCD)

Consider ℓ_1 steepest descent $\equiv$ Greedy Coordinate Descent (GCD) for Logistic Regression

Greedy Coordinate Descent for Logistic Regression

Greedy Coordinate Descent for Logistic Regression

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

At iteration $k \geq 0$:

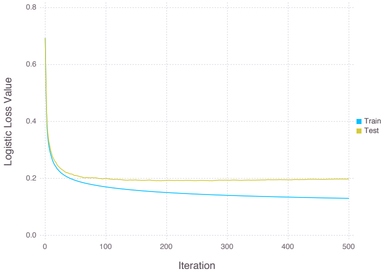
1. Compute $\nabla L_n(\beta^k)$
2. Compute $j_k \in \arg \max_{j \in \{1,\ldots,p\}} |\nabla L_n(\beta^k)_j|$
3. Set $\beta^{k+1} \leftarrow \beta^k - \alpha_k \mathrm{sgn}(\nabla L_n(\beta^k)_{j_k}) e_{j_k}$

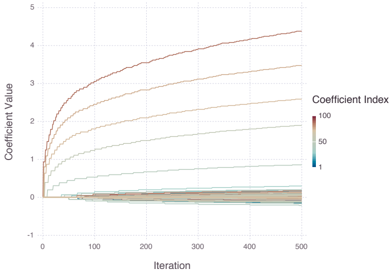
Why use Greedy Coordinate Descent for Logistic Regression?

- 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)

Implicit Regularization and Variable Selection Properties

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





Compare with explicit regularization schemes (ℓ_1, ℓ_2, etc.)

How Can SDGN Inform Logistic Regression?

Some questions:

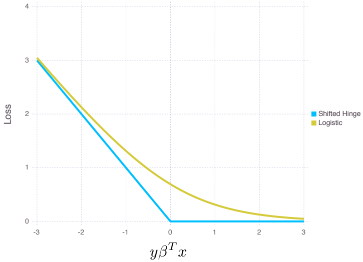
- How do the computational guarantees for the Steepest Descent family specialize to the case of Logistic Regression?
- What role does problem structure/conditioning play in these guarantees?
- Can we say anything further about the convergence properties of the Steepest Descent family in the special case of Logistic Regression?

Elementary Properties of the Logistic Loss Function

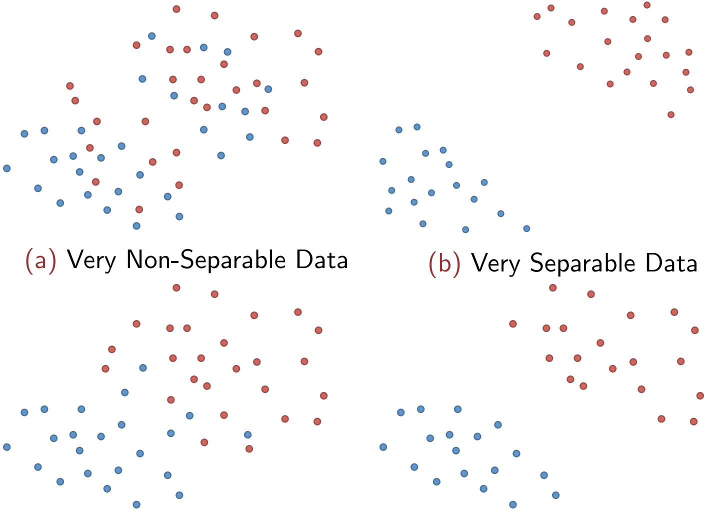
$$L_n^* \quad := \quad \min_{\beta} \quad L_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} \ln(1+\exp(-y_i\beta^T x_i))$$

Logistic regression "ideally" seeks β 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: Non-Separable and Separable Data



(a) Very Non-Separable Data

(b) Very Separable Data

(c) Mildly Non-Separable Data

(d) Mildly Separable Data

Separable and Non-Separable Data

Separable Data

The data is separable if there exists $\bar{\beta}$ for which

$$y_i \cdot (\bar{\beta})^T x_i > 0 \quad \text{for all } i = 1, \ldots, n$$

Non-Separable Data

The data is non-separable if it is not separable, namely, every β satisfies

$$y_i \cdot (\beta)^T x_i \leq 0 \quad \text{for some } i \in \{1, \ldots, n\}$$

Separable Data

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

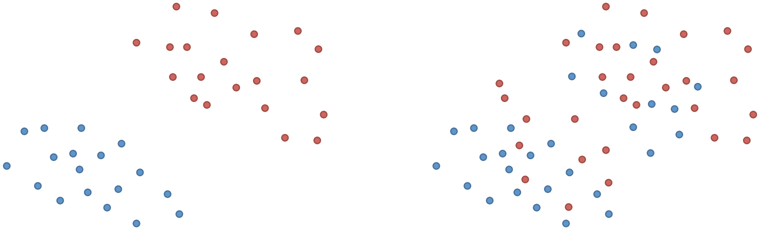
The data is separable if there exists $\bar{\beta}$ for which

$$y_i \cdot (\bar{\beta})^T x_i > 0 \quad \text{for all } i = 1, \ldots, n$$

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 classifier ....

Separable and Non-Separable Data



(a) Separable

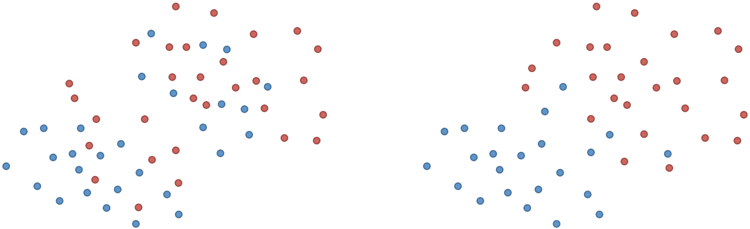
(b) Non-Separable

Results in the Non-Separable Case

Results in the Non-Separable Case

Non-Separable Data and Problem Behavior/Conditioning

Let us quantify the degree of non-separability of the data.



(a) Very non-separable data

(b) Mildly non-separable data

We will relate this to problem behavior/conditioning....

Non-Separability Condition Number $\mathrm{DegNSEP}^*$

Definition of Non-Separability Condition Number $\mathrm{DegNSEP}^*$

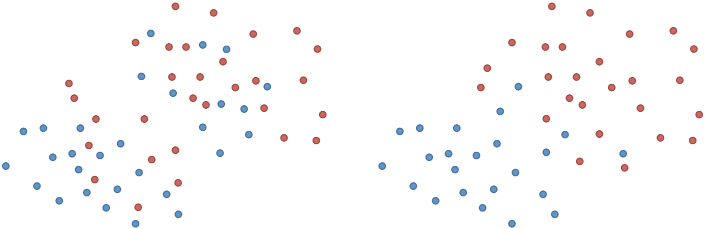
$$\begin{array}{rcc} \mathrm{DegNSEP}^* := & \min\limits_{\beta \in \mathbb{R}^p} & \frac{1}{n}\sum_{i=1}^n [y_i \beta^T x_i]^- \\ & \text{s.t.} & \|\beta\| = 1 \end{array}$$

$\mathrm{DegNSEP}^*$ is the least average misclassification error (over all normalized classifiers)

$\mathrm{DegNSEP}^* > 0$ if and only if the data is strictly non-separable

Non-Separability Measure $\mathrm{DegNSEP}^*$

$$\begin{aligned} \mathrm{DegNSEP}^* := \quad \min_{\beta \in \mathbb{R}^p} \quad & \tfrac{1}{n} \textstyle\sum_{i=1}^n [y_i \beta^T x_i]^- \\ \text{s.t.} \quad & \|\beta\| = 1 \end{aligned}$$



(a) $\mathrm{DegNSEP}^*$ is large

(b) $\mathrm{DegNSEP}^*$ is small

DegNSEP* and Problem Behavior/Conditioning

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

$$\begin{array}{rcl} \mathrm{DegNSEP}^* := & \displaystyle\min_{\beta\in\mathbb{R}^p} & \frac{1}{n}\sum_{i=1}^n [y_i\beta^T x_i]^- \\ & \text{s.t.} & \|\beta\| = 1 \end{array}$$

Theorem: Non-Separability and Sizes of Optimal Solutions

Suppose that the data is non-separable and $\mathrm{DegNSEP}^* > 0$. Then

1. the logistic regression problem LR attains its optimum,
2. for every optimal solution β^* of LR it holds that
$$\|\beta^*\| \le \frac{L_n^*}{\mathrm{DegNSEP}^*} \le \frac{\ln(2)}{\mathrm{DegNSEP}^*}$$, and
3. for any β it holds that $\|\beta\| \le \dfrac{L_n(\beta)}{\mathrm{DegNSEP}^*}$.

Computational Guarantees for Steepest Descent family: Non-Separable Case

Theorem: Computational Guarantees for Steepest Descent family: Non-Separable Case

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

(i) (training error): $L_n(\beta^k) - L_n^* \ \leq \ \frac{2(\ln(2))^2\|\mathbf{X}\|_{\cdot,2}^2}{k \cdot n \cdot (\mathrm{DegNSEP}^*)^2}$

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

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

Computational Guarantees for Steepest Descent family: Non-Separable Case

Theorem: Computational Guarantees for Steepest Descent family: Non-Separable Case

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

(i) (training error): $L_n(\beta^k) - L_n^* \ \leq \ \frac{2(\ln(2))^2\|\mathbf{X}\|_{\cdot,2}^2}{k\cdot n\cdot(\mathrm{DegNSEP}^*)^2}$

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

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

Reaching Linear Convergence

Reaching Linear Convergence using Steepest Descent with a Given Norm for Logistic Regression

For logistic regression, does SDGN exhibit linear convergence?

Some Definitions/Notation

Definitions:

- $R := \max_{i \in \{1,\ldots,n\}} \|x_i\|_2$ (maximum ℓ_2 norm of the feature vectors)
- $H(\beta^*)$ denotes the Hessian of $L_n(\cdot)$ at an optimal solution β^*
- $\lambda_{\mathrm{pmin}}(H(\beta^*))$ denotes the smallest non-zero (and hence positive) eigenvalue of $H(\beta^*)$
- NormRatio $:= \max_{\beta \neq 0} \|\beta\|/\|\beta\|_2$

Reaching Linear Convergence of Steepest Descent family for Logistic Regression

Theorem: Reaching Linear Convergence of Steepest Descent family for Logistic Regression

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

$$\check{k} := \frac{16\ln(2)^2\|\mathbf{X}\|_{\cdot,2}^4 R^2(\mathrm{NormRatio})^2}{9n^2(\mathrm{DegNSEP}^*)^2\lambda_{\mathrm{pmin}}(H(\beta^*))^2} \ .$$

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

$$L_n(\beta^k) - L_n^* \leq (L_n(\beta^{\check{k}}) - L_n^*)\left(1 - \frac{\lambda_{\mathrm{pmin}}(H(\beta^*))n}{\|\mathbf{X}\|_{\cdot,2}^2(\mathrm{NormRatio})^2}\right)^{k-\check{k}} \ .$$

Reaching Linear Convergence of Steepest Descent family for Logistic Regression, cont.

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:

$$\lambda_{\mathrm{pmin}}(H(\beta^*)) \;\geq\; \tfrac{1}{4n}\lambda_{\mathrm{pmin}}(\mathbf{X}^T\mathbf{X}) \exp\left(-\frac{\ln(2)\|\mathbf{X}\|_{\cdot,\infty}}{\mathrm{DegNSEP}^*}\right)$$

- 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 . . .)

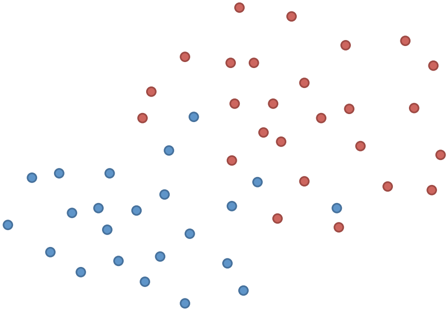
DegNSEP* and "Perturbation to Separability"

$$\begin{aligned} \mathrm{DegNSEP}^* := \quad \min_{\beta \in \mathbb{R}^p} \quad & \frac{1}{n}\sum_{i=1}^n [y_i \beta^T x_i]^- \\ \text{s.t.} \quad & \|\beta\| = 1 \end{aligned}$$

Theorem: DegNSEP* is the "Perturbation to Separability"

$$\begin{aligned} \mathrm{DegNSEP}^* = \quad \inf_{\Delta x_1, \ldots, \Delta x_n} \quad & \frac{1}{n}\sum_{i=1}^n \|\Delta x_i\|_* \\ \text{s.t.} \quad & (x_i + \Delta x_i, y_i), i = 1, \ldots, n \text{ are separable} \end{aligned}$$

Illustration of Perturbation to Separability



Results for Some other Methods

Results for Some other Methods

Standard Accelerated Gradient Method (AGM)

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

Lipschitz gradient: $\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2$ for all $x, y \in \mathbb{R}^p$

Accelerated Gradient Method (AGM)

Given $x^0 \in \mathbb{R}^p$ and $z^0 := x^0$, and $i \leftarrow 0$. Define step-size parameters $\theta_i \in (0, 1]$ recursively by $\theta_0 := 1$ and θ_{i+1} satisfies $\frac{1}{\theta_{i+1}^2} - \frac{1}{\theta_{i+1}} = \frac{1}{\theta_i^2}$.

At iteration k:

1. Update : $y^k \leftarrow (1 - \theta_k)x^k + \theta_k z^k$

   $x^{k+1} \leftarrow y^k - \frac{1}{L}\nabla f(y^k)$

   $z^{k+1} \leftarrow z^k + \frac{1}{\theta_k}(x^{k+1} - y^k)$

Computational Guarantees for Accelerated Gradient Method (AGM) for Logistic Regression

Theorem: Computational Guarantees for Accelerated Gradient Method (AGM) for Logistic Regression

Consider the AGM applied to the Logistic Regression problem initiated at $\beta^0 := 0$, and suppose that the data is non-separable. Then for each $k \geq 0$ it holds that:

$$\text{(training error):} \qquad L_n(\beta^k) - L_n^* \;\leq\; \frac{2(\ln(2))^2 \|\mathbf{X}\|_{2,2}^2}{n \cdot (k+1)^2 \cdot (\mathrm{DegNSEP}^*)^2}$$

AGM with Simple Re-Starting (AGM-SRS)

Assume that $0 < F^* := \ \text{minimum}_x \ F(x)$

Accelerated Gradient Method with Simple Re-Starting (AGM-SRS)

```
Initialize with $x^0 \in \mathbb{R}^p$ .
Set $x_{1,0} \leftarrow x^0$ , $i \leftarrow 1$ .

At outer iteration $i$:

  1 **Initialize inner iteration.** $j \leftarrow 0$
  2 **Run inner iterations.** At inner iteration $j$:
        If $\frac{F(x_{i,j})}{F(x_{i,0})} \geq 0.8$ , then:
              $x_{i,j+1} \leftarrow \text{AGM}(F(\cdot),\ x_{i,0},\ j+1)$ ,
              $j \leftarrow j+1$, and Goto step 2.
        Else $x_{i+1,0} \leftarrow x_{i,j}$, $i \leftarrow i+1$, and Goto step 1.
```

"$x_{i,j} \leftarrow \text{AGM}(F(\cdot),\ x_{i,0},\ j)$" denotes assigning to $x_{i,j}$ the j^{th} iterate of AGM applied with objective function $F(\cdot)$ using the initial point $x_{i,0} \in \mathbb{R}^p$

Computational Guarantee for AGM with Simple Re-Starting for Logistic Regression

Computational Guarantee for Accelerated Gradient Method with Simple Re-Starting for Logistic Regression

Consider the AGM with Simple Re-Starting applied to the Logistic Regression problem initiated at $\beta^0 := 0$, and suppose that the data is non-separable. Within a total number of computed iterates k that does not exceed

$$\frac{5.8\|\mathbf{X}\|_{2,2}}{\sqrt{n}\cdot \mathrm{DegNSEP}^*} + \frac{8.4\|\mathbf{X}\|_{2,2}\cdot L_n^*}{\sqrt{n}\cdot \mathrm{DegNSEP}^*\cdot\sqrt{\varepsilon}} ,$$

the algorithm will deliver an iterate β^k for which

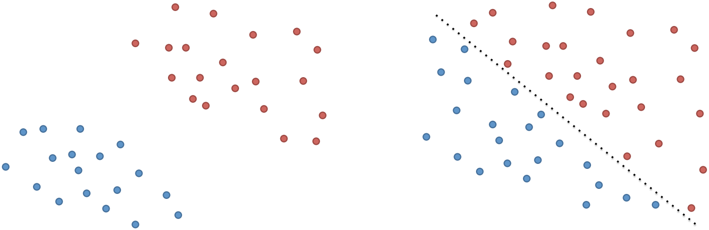
$$L_n(\beta^k) - L_n^* \ \le \ \varepsilon \ .$$

Results in the Separable Case

Results in the Separable Case

Separable Data and Problem Behavior/Conditioning

Let us quantify the degree of separability of the data.



(a) Very separable data

(b) Barely separable data

We will relate this to problem behavior/conditioning....

Separability Condition Number DegSEP^*

Definition of Non-Separability Condition Number DegSEP^*

$$\begin{array}{rcl} \mathrm{DegSEP}^* := & \max_{\beta \in \mathbb{R}^p} & \min_{i \in \{1,\ldots,n\}} [y_i \beta^T x_i] \\ & \text{s.t.} & \|\beta\| \leq 1 \end{array}$$

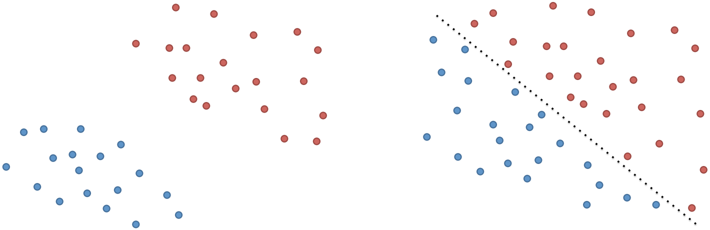
DegSEP^* maximizes the minimal classification value $[y_i \beta^T x_i]$ (over all normalized classifiers)

DegSEP^* is simply the "maximum margin" in machine learning parlance

$\mathrm{DegSEP}^* > 0$ if and only if the data is separable

Separability Measure DegSEP*

$$
\begin{aligned}
\text{DegSEP}^* := \quad & \max_{\beta \in \mathbb{R}^p} \quad \min_{i \in \{1,\ldots,n\}} [y_i \beta^T x_i] \\
\text{s.t.} \quad & \|\beta\| \le 1
\end{aligned}
$$



(a) DegSEP* is large

(b) DegSEP* is small

DegNSEP* and Problem Behavior/Conditioning

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

$$\begin{array}{rcl} \mathrm{DegSEP}^* := & \displaystyle\max_{\beta\in\mathbb{R}^p} & \displaystyle\min_{i\in\{1,\ldots,n\}} [y_i\beta^T x_i] \\ & \text{s.t.} & \|\beta\| \le 1 \end{array}$$

Theorem: Separability and Non-Attainment

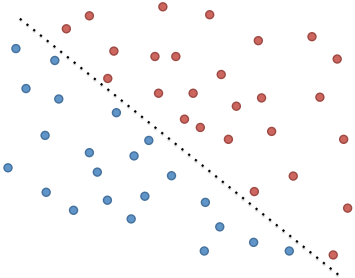
Suppose that the data is separable. Then $\mathrm{DegSEP}^* > 0$, $L_n^* = 0$, and LR does not attain its optimum.

Despite this, it turns out that the Steepest Descent family is reasonably effective at finding an approximate margin maximizer as we shall shortly see....

Margin function $\rho(\beta)$

Margin function $\rho(\beta)$

$$\rho(\beta) := \min_{i \in \{1,\dots,n\}} [y_i \beta^T x_i]$$



Computational Guarantees for Steepest Descent family: Separable Case

Theorem: Computational Guarantees for Steepest Descent family: Separable Case

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

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

$$\rho(\bar{\beta}^i) \geq \frac{.18 \cdot \mathrm{DegSEP}^*}{n} .$$

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

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

Computational Guarantees for Steepest Descent family: Separable Case

Theorem: Computational Guarantees for Steepest Descent family: Separable Case

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

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

$$\rho(\bar{\beta}^i) \;\geq\; \frac{.18 \cdot \mathrm{DegSEP}^*}{n} \; .$$

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

(iii) (gradient norm): $\min\limits_{i \in \{0,\ldots,k\}} \|\nabla L_n(\beta^i)\|_* \;\leq\; \|\mathbf{X}\|_{\cdot,2} \sqrt{\frac{\ln(2)}{2n \cdot (k+1)}}$

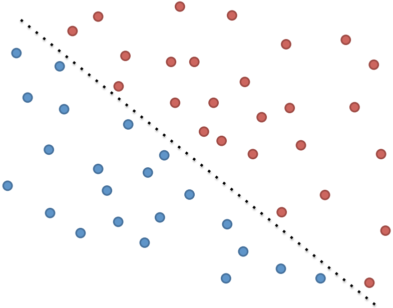
DegSEP* and "Perturbation to Non-Separability"

$$\begin{aligned} \mathrm{DegSEP}^* := \quad \max_{\beta \in \mathbb{R}^p} \quad & \min_{i \in \{1,\ldots,n\}} [y_i \beta^T x_i] \\ \text{s.t.} \quad & \|\beta\| \leq 1 \end{aligned}$$

Theorem: DegSEP* is the "Perturbation to Non-Separability"

$$\begin{aligned} \mathrm{DegSEP}^* = \quad \inf_{\Delta x_1,\ldots,\Delta x_n} \quad & \max_{i \in \{1,\ldots,n\}} \|\Delta x_i\|_* \\ \text{s.t.} \quad & (x_i + \Delta x_i, y_i),\, i = 1, \ldots, n \text{ are non-separable} \end{aligned}$$

Illustration of Perturbation to Non-Separability



Other Issues

Some other topics not mentioned (still ongoing):

- Other first-order methods for logistic regression (accelerated gradient descent, randomized methods, etc.
- high-dimensional regime $p > n$, define $\mathrm{DegNSEP}_k^*$ and DegSEP_k^* for restricting β to satisfy $\|\beta\|_0 \leq k$
- Numerical experiments comparing methods
- Other...

Summary

- Some old and new results for Steepest Descent in a Given Norm (SDGN)

- Analyizing SDGN for Logistic Regression: separable/non-separable cases

- Non-Separable case
  - condition number $\mathrm{DegNSEP}^*$
  - computational guarantees for SGDN including reaching linear convergence

- Separable case
  - condition number DegSEP^*
  - computational guarantees for SGDN including computing an approximate maximum margin classifier