Case Study: Logistic Regression

Ruoyu Sun

Review Questions

- What is the key factor of the convergence rate of GD, for strongly convex functions?

- What is the definition of this factor?

- If the function is convex but not strongly convex, then what does the convergence rate depend on?

- What if the function is non-convex? How to estimate the convergence speed of GD?

Today

- **Today**: (Theory-Guided) Case study of logistic regression (LR), mostly using what you have learned
  - Note: examples may be (much) more complicated than theory
- After today's course, you will be able to
  - use a few standard tools for analyzing algorithms
  - describe the convergence result of logistic regression
  - explain the algorithm behavior of low-dim LR using existing theory

Review of Theories You Learned

- Optimality conditions: check local-min/global-min
    - Convexity is crucial
- When does GD “converge”? (every limit point is stationary)
    - Stepsize choice: $< 2/L$ with L-Lipschitz gradient
    - line search
- Convergence rate? depend on κ for strongly convex problems

Apply Theory?

- We will see: it is nontrivial to apply these theories to even a simple problem
- Image yourself being an data scientist/engineer/etc., trying to solve a problem.
  - Step 1: specify the problem
  - Step 2: write or modify code to solve it
  - Step 3: Observe algorithm behavior, good or bad? expected or not?
  - Step 4: If not good/expected, go to Step 2
- Keep in mind: as an optimizer, your goal is to understand the behavior (as much as possible)
  - The goal of a regular data scientist is probably “making it work”

Outline

Warmup: 1-dim Linear Regression

1-dim logistic regression

- Logistic Regression Introduction
- First Data Setting, Stopping Criteria
- Second Data Setting

Formal Results for Logistic Regression

Experiments of Higher-Dim LR

Warm-up: 1-dim linear regression (basketball problem)

Before discussing logistic regression, we do a warm-up for linear regression.

- Set a baseline for the behavior
- Try to understand everything you observe

Problem setup: $\min_w (p - cw)^2$.

- Problem parameters: $p = 1, c = 5$.

Algorithm: $w^+ = w - \alpha \nabla f(w)$.

- Initialization: $w^0 = 5$. (or randn(1,1); let's do fixed first)
- Max-iteration: MaxIte = 50. (can change)

Tuning hyperparameter: α.

1-dim regression: threshold

Start from

- $\alpha = 1$, diverge
- $\alpha = 0.1$, diverge
- $\alpha = 0.01$, converge.

You want to know the threshold, and find the threshold is 0.08.

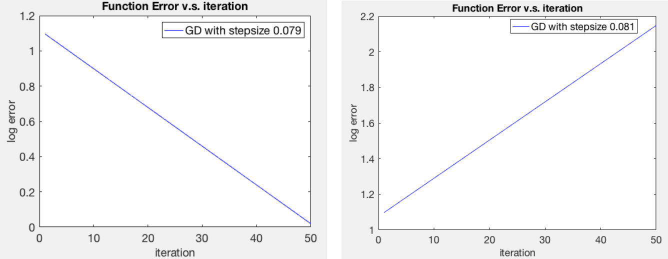


Figure: Left: stepsize 0.081; Right: stepsize 0.079

By theory, the threshold of $\alpha = 2/L = 2/c^2 = 0.08$.

1-dim regression: speed

You want to find α so it converges fast.

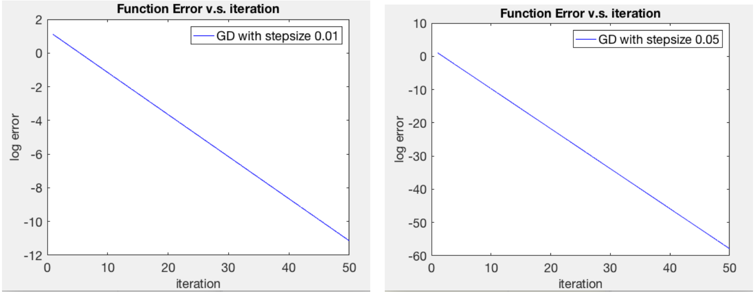


Figure: Left: stepsize 0.01; Right: stepsize 0.05

Stepsize affects the convergence speed a lot.

$\alpha = 0.01$, to achieve 10^{-10}, need 40 iterations.

$\alpha = 0.05$, to achieve 10^{-10}, need 10 iterations – faster

What is the optimal stepsize? $2/(L+\mu) = 1/L = 0.04$.

1-dim regression: summary

Summary of 1-dim linear regression: stepsize affects convergence speed.

Convergence or not, two phases:

- Phase 1: $\alpha > 0.08 = 2/L$. Diverge.
- Phase 2: $0 < \alpha < 0.08 = 2/L$. Converge.

Convergence speed, two sub-phases:

- Sub-Phase 1: $0 < \alpha < 0.04 = 2/(L + \mu)$ bigger stepsize implies faster
- Sub-Phase 2: $2/(L + \mu) = 0.04 < \alpha < 0.08 = 2/L$, bigger stepsize imples slower.

Outline

Warmup: 1-dim Linear Regression

1-dim logistic regression
- Logistic Regression Introduction
- First Data Setting, Stopping Criteria
- Second Data Setting

Formal Results for Logistic Regression

Experiments of Higher-Dim LR

Classification Problem

- Quadratic functions are typical but also special.
- For understanding optimization, let's check a non-quadratic convex problem: logistic regression.
- Classification: one of the key problems of machine learning
- Example: classify cat or dog.

Linear Classification

Classification: $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, where $y_i \in \{1, -1\}$.

- Find $\mathbf{w}$ such that dist$(\mathbf{w}^T\mathbf{x}_i, y_i)$ is small.
- Choose a loss function $\log(1 + \exp(-y\hat{y}))$

$$\min_{w} \sum_i \text{dist}(\mathbf{w}^T\mathbf{x}_i, y_i) = \sum_i \log(1 + \exp(-y_i\mathbf{w}^T\mathbf{x}_i)).$$

Side: Popularity of Logistic Regression (LR)

Google search ”logistic regression”: about 10 million results.

“There is a perception that LR is slow, unstable, and unsuitable for large learning or classification tasks” –Komarak’04 ”Logistic Regression for Data Mining and High- Dimensional Classification”

But It is still very popular nowadays:

- deep learning is built on LR
- very commonly used in IT companies like Facebook, Google, etc.

There are many tutorials on LR.

- One is Stanford CS231n course webpage
- Or you can check a detailed tutorial at http://www.dataschool.io/guide-to-logistic-regression/

We focus on optimization issues in this course.

Simplest problem: 1-dim case

- **Problem setup**:
  - Two data points in $\mathbb{R}$, x_1 and x_2.
  - Labels $y_1 = 1$, $y_2 = -1$.

- Remark: typically, one would use $wx_i + b$ to classify; but for simplicity, let's just use wx_i for now.

- **Objective function**

$$f(w) = \frac{1}{2}\log(1+\exp(-wx_1)) + \frac{1}{2}\log(1+\exp(wx_2)).$$

Algorithm: $w^+ = w - \alpha\nabla f(w)$.

- Initialization: $w_0 = 5$. (or randn(1,1); let's use a fixed w_0 first)
- Max-iteration: MaxIte = 20. (can change)

Tuning hyperparameter: α.

Preliminary Analysis

Before running the algorithm, let's do some simple calculations.

- **First**, is the problem convex?
- Obtain $f''(w) = x_1^2 \frac{e^{wx_1}}{(1+e^{wx_1})^2} + x_2^2 \frac{e^{wx_2}}{(1+e^{wx_2})^2} > 0$.
  So it is strictly convex.
- **Second**, is the problem L-smooth?

  Since $\frac{e^{wx_i}}{(1+e^{wx_i})^2} \leq \frac{e^{wx_i}}{1+e^{wx_i}} \leq 1$, we have $f''(w) \leq x_1^2 + x_2^2$.

  Thus f is L-smooth with $L = x_1^2 + x_2^2$.
- **Other**: is the function lower bounded?
  Yes! Lower bounded by 0.

“First experiment”

Data Setting 1: Set $x_1 = 1, x_2 = 2$.

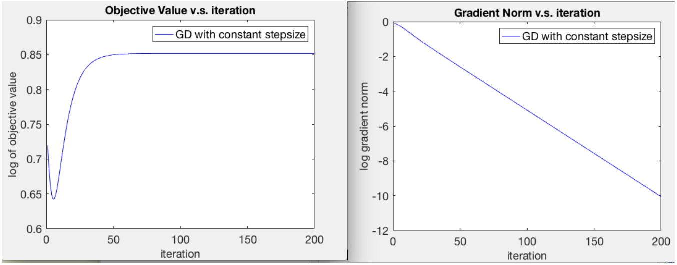


Figure: Function Value. $\alpha = 0.2$ for 1-dim logistic regression

- Why increasing?

  BUG! Used $\log(1 + \exp(y_i w^T x_i))$, missing a “ $-$ ”.

- Seriously? Why do you show a bug?
  I’ll explain later.

First experiment

Data Setting 1: Set $x_1 = 1, x_2 = 2$.

Pick $\alpha = 0.2$. Figure of the algorithm for 200 iterations:

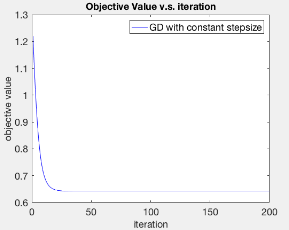


Figure: Function Value. $\alpha = 0.2$ for 1-dim logistic regression

Question: Has it already converged?

Issue: in theory, use $f(x^r) - f^*$ to measure error,

but in practice don't know f^*

Performance Metric and Stopping Criteria

One choice: $f(x^r) - f(x^{r-1})$;

last few iterations of $f(x^r)$: $0.6420, 0.6420, 0.6420, \ldots$

Drawbacks:

- What's the point? Only implies $f(x^r)$ converges (Type-1); not very meaningful

A better choice: gradient norm $\|\nabla f(x^r)\|$

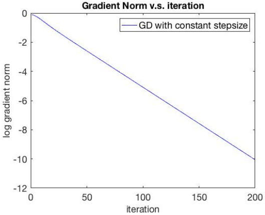


Figure: Gradient norm. $\alpha = 0.2$ for 1-dim logistic regression

Stopping Criteria

Two metric: $\nabla f(x^r)$ and $|f(x^r) - f(x^{r+1})|$.

Can stop the algorithm when either of the three holds:

- ST1: when $\|\nabla f(x^r)\| \leq \epsilon$
- ST2: when $|f(x^r) - f(x^{r+1})| \leq \epsilon$
- ST3: when $\|x^r - x^{r+1}\| \leq \epsilon$
  - For GD with constant stepsize, ST3 equivalent to ST1

Typical choice of error: $\epsilon = 10^{-3}$ (low precision), or $\epsilon = 10^{-10}$ (high precision)

Practice: relation between the three criteria?

Second experiment

Data Setting 2: Set $x_1 = 1, x_2 = -2$.

Pick $\alpha = 0.2$. Figure of the algorithm for 200 iterations (use log scale):

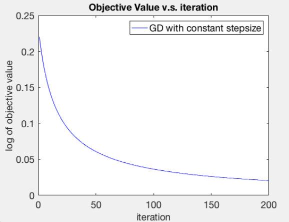


Figure: Function Value. $\alpha = 0.2$ for 1-dim logistic regression

Question: Has it “converged”?

Converged or Not?

Let’s run 20k iterations. That should be enough, uhh?

Check log of function values, and gradient norms

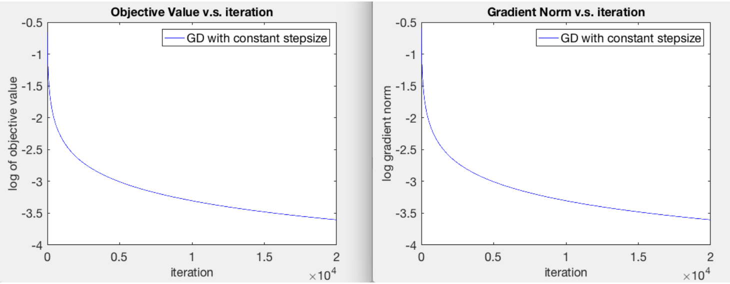


Figure: Left: Log of Function Value. Right: Gradient Norm. $20,000$ iterations

Again, has it “converged”? Or, should we count it as “converged”?

- Yes, since the loss function is small?
- No, since the gradient norm is still not small enough?

Discussion: Why This Happens

So strange... In all previous cases (linear regression, or Data Setting 1), either diverge or converge in the sense that “error” $< 10^{-10}$.

Re-examine the theory

- If $\alpha < 2/L$, then every limit point is stationary point.
- If convex, then every stationary point is globally-optimal.

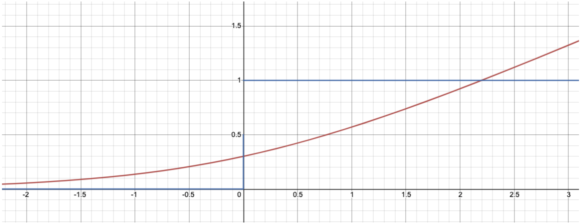
So, it should finally converge to a stationary point, which is globally optimal (right?). Maybe 20,000 iterations is not enough?

First Possible Issue

Theory says... every limit point is stationary (global-min for convex)

First Possible issue: Non-existence of global-min.

- Recall the graph of $\log(1 + \exp(-z))$, where $z = y\hat{y}$.



- **Observation**: Although strictly convex and lower bounded, it has no stationary point.

- Iterates diverge: $\|w^k\|$ goes to ∞ to minimize $f(w)$

- No limit point... so we have nothing to say about "convergence"?

First Issue is Not Really an Issue

- Even though “every limit point is stationary” is meaningless in this problem, another conclusion “ $\|\nabla f(x^k)\| \to 0$ ” is always meaningful (since there is lower bound).

- **Claim:** Using GD with constant stepsize $\alpha < 2/L$ for this 1-diim LR problem, we have $\|\nabla f(w^r)\| \to 0.$
  - **Implication**: If we run GD for infinitely many iterations, $\|\nabla f(w^r)\|$ will fall below 10^{-10}

- We re-affirm: the theory is still meaningful;
  but how to explain the practice “after 20k iterations, GD is still not fully convergent”?

Second Issue

- **Second Possible Issue**: Convergence speed.
- Now we know $f(x^r)$ does converge to 0. But why is it so slow?
- Two possibilities: (A) Another fundamental reason; (B) Bug.

  Some of you may say: I'm a programming genius, there cannot be bug!

  But for most people, the possibility of "bug" is hard to eliminate (especially for unfamiliar subject)
- Common debug trick: check simple case; hopefully the behavior for the simple case is expected

  But now, you are dealing with 2 data points, 1-dim, convex problem.

  You still need to know what should be expected (the fundamental reason)

Second Issue: convergence speed

- A first thought: condition number too large.
- Compute condition number at every point, turn out to be... 1.
  - 1-dim problem, certainly at each point $\lambda_{\max} = \lambda_{\min}$
- What's wrong?

Issue due to condition number?

We compute $L(w^r)$ and $\mu(w^r)$ at all iterates:

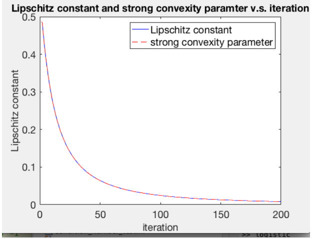


Figure: L and μ of various iterates

Observation: $L(w^r) = \mu(w^r)$ is decreasing.

Definition of condition number: $\mu \leq f''(w) \leq L$ for all w, then $\kappa = \frac{L}{\mu}$.

E.g. $L(w^1) = 0.5$, $L(w^{200}) \approx 0.008$, then κ is at least $0.5/0.008 = 62.5$.

Result for non-strongly convex functions

Theoretically speaking, the condition number of f is? ∞

- Reason: $f''(w) \to 0$ as $|w| \to \infty$, thus the best μ satisfying $\mu \leq f''(w)$ is $\mu = 0$

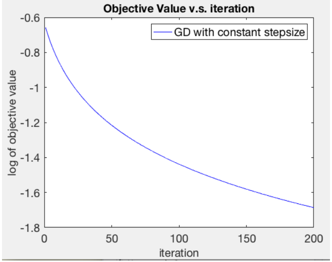
So, the convergence rate result on κ does not apply here!

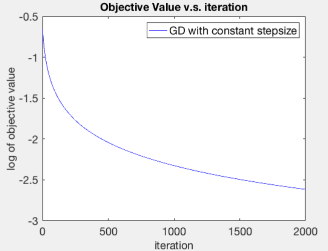
Then can we have convergence speed guarantee?

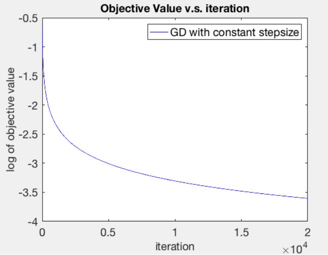
- Recall: for convex functions, GD converges at rate $O(1/K)$
- There is a caveat; cannot directly apply; see formal result in next section

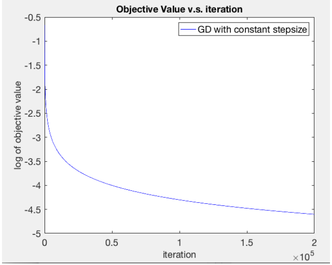
Verifying Sublinear Rate in Practice

Run for 200, 2k, 20k, 200k iterations.









Verifying Sublinear Rate in Practice (cont'd)

Collect the simulation results in a table:

$\#$ of Iterations	200	2k	20k	200k
error	$10^{-1.6}$	$10^{-2.6}$	$10^{-3.6}$	$10^{-4.6}$

Table: Error v.s. Iterations

- **Observation**: To reduce error to 1/10, need 10 times more iterations; behaves like $O(1/K)$.
- So the performance (almost) matches the theory! (finally)

Why Two Data Settings Different

Why are Data Setting 1 $(1, 2)$ and Setting 2 $(1, -2)$ so different?

What if we use $(1, 5)$? or $(-0.3, -6)$?

Two typical data settings:

- Separable case
- Non-separable case
- In separable case, w diverges, arrive at flat region!
- In non-separable case, w stays bounded, so essentially“strongly convex”!

Question: Why care about divergence of iterates? (objective small is enough)

- Because it may affect convergence rate a lot!

Summary of Observations for 1-dim LR

- Since $\log(1+e^{-z})$ is strictly convex, but NOT strongly convex, the worst-case convergence rate shall be sublinear, such as $O(1/K)$.
- We observe two convergence patterns:
  - Separable case: sublinear rate
  - Non-separable case: usually linear rate

Outline

Warmup: 1-dim Linear Regression

1-dim logistic regression

- Logistic Regression Introduction
- First Data Setting, Stopping Criteria
- Second Data Setting

Formal Results for Logistic Regression

Experiments of Higher-Dim LR

Summary of Results in General Logistic Regression

We state a few (plausible) results for general LR.

Result 1: If the data are **linearly separable**, then:

- **Result 1a:** the objective function is not strongly convex; in addition, there is no global minimum.
- **Result 1b:** GD with proper stepsize converges to global infimum at a sublinear rate.

Result 2: If the data are **not linearly separable**, then:

- **Result 2a:** the objective is coercive; but still non-strongly convex; it has a unique global-min
- **Result 2b** (without proof presented): GD with proper stepsize converges to the unique global-min at a linear rate.

Setting and Basic Properties

Data set $\{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$.

Optimization problem:

$$\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) \triangleq \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})).$$

Claim 1: $F(\mathbf{w})$ is L-smooth, where $L = \frac{1}{4}\lambda_{\max}(X^\top X)$. (see homework)

Claim 2: $F(\mathbf{w})$ is convex.

Proof: The Hessian at $\mathbf{w}$ is (see homework)

$$\nabla^2 f(\mathbf{w}) = \sum_{i=1}^n \frac{e^{-y_i w^\top x_i}}{(1 + e^{-y_i w^\top x_i})^2} x_i x_i^\top.$$

Since $\frac{e^{-y_i w^\top x_i}}{(1+e^{-y_i w^\top x_i})^2} \geq 0$, and $x_i x_i^\top$ is PSD, each term is PSD; then the sum is also PSD. □

Attempt in Applying Existing Results

- Convergence rate for a convex L-smooth function f?

- **Prop 1a** (**Prop 2 in Lec 1c**) for any optimal solution $\mathbf{w}^*$,

$$f(\mathbf{w}^r) - f(\mathbf{w}^*) \leq \frac{2L}{r} \|\mathbf{w}^0 - \mathbf{w}^*\|^2.$$

- Wait...something is wrong!

  For logistic regression, global-min may not exist! (global infimum exists)

  - Cannot apply Prop 1a directly;
  - Prop 1a is still correct; "for any optimal solution, xx holds" does not imply "optimal solution exists"!

Convergence Result for LR: Linearly Separable Case

A variant of Prop 2 in Lec 1c can be useful for “no global-min” case.

Proposition 1b (extension of Prop 1a): Suppose f is a convex and L-smooth. Consider $\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{w})$. Suppose GD with stepsize $1/L$ generates $\{\mathbf{w}^r\}$, then for any $\hat{\mathbf{w}}$,

$$f(\mathbf{w}^r) - f(\hat{\mathbf{w}}) \leq \frac{2L}{r-1}\|\mathbf{w}^0 - \hat{\mathbf{w}}\|^2.$$

- Remark: The result applies to any fixed $\hat{w}$, not just optimal solutions.

Assumption 1 (linearly separable with margin γ): There exists a $\mathbf{w}^*$ which satisfies $\|\mathbf{w}^*\| = 1$ and $y_i \mathbf{x}_i^\top \mathbf{w}^* \geq \gamma > 0, \forall\, i = 1, 2, \ldots, n$.

Theorem 1 (Convergence rate of GD for LR on linearly separable case) Suppose F is L-smooth and stepsize $\alpha = 1/L$. Under Assumption 1,

$$F(\mathbf{w}^{K+1}) \leq \frac{1 + 4L\ln(Kn\gamma^2)^2}{K\gamma^2} + \frac{4L\|\mathbf{w}^0\|^2}{K} = \Theta(\frac{\ln(K)}{K}),\ \forall K.$$

Proof of Prop 1b (reading)

Step 1 (difference of function value):

$$F(w_{r+1}) - F(w_r) \le \langle \nabla F(w_r), w_{r+1} - w_r \rangle + \frac{L}{2}\|w_{r+1} - w_r\|^2 \tag{1a}$$

$$= -\frac{1}{L}\|\nabla F(w_r)\|^2 + \frac{L}{2}\|w_{r+1} - w_r\|^2 \le -\frac{1}{2L}\|\nabla F(w_r)\|^2. \tag{1b}$$

where (1a) is by the L-smoothness of F (by the descent lemma); (1b) is due to the identity $w_{r+1} - w_r = -\frac{1}{L}\nabla f(w_r)$. This implies $F(w_r)$ is strictly decreasing over r.

Step 2 (difference of iterate error): By convexity, $F(\hat{w}) \ge F(w_r) + \langle \nabla F(w_r), \hat{w} - w_r \rangle$. Then

$$\|w_{r+1} - \hat{w}\|^2 - \|w_r - \hat{w}\|^2 = \|w_r - \eta\nabla F(w_r) - \hat{w}\|^2 - \|w_r - \hat{w}\|^2 \tag{2a}$$

$$= \eta^2\|\nabla F(w_r)\|^2 - 2\eta\langle \nabla F(w_r), w_r - \hat{w} \rangle \tag{2b}$$

$$\overset{(i)}{\le} \frac{1}{L^2}\|\nabla F(w_r)\|^2 + \frac{2}{L}(F(\hat{w}) - F(w_r)). \tag{2c}$$

$$\overset{(ii)}{\le} \frac{2}{L}(F(w_r) - F(w_{r+1})) + \frac{2}{L}(F(\hat{w}) - F(w_r)) \tag{2d}$$

$$= \frac{2}{L}(F(\hat{w}) - F(w_{r+1})) \tag{2e}$$

Step 3: Telescope sum. We obtain

$$\|w_{r+1} - \hat{w}\|^2 - \|w_0 - \hat{w}\|^2 \le \frac{2}{L}\sum_{i=1}^{r+1}(F(\hat{w}) - F(w_i)) \le \frac{2}{L}r(F(\hat{w}) - F(w_{r+1})). \tag{3}$$

Thus $F(w_{r+1}) - F(\hat{w}) \le \frac{L}{2r}\|w_0 - \hat{w}\|^2$.

Proof of Theorem 1 (reading)

For a given K, we pick a reference point $\mathbf{v}$ (to be decided later), then by Prop. 1b,

$$F(\mathbf{w}^{K+1}) \leq F(\mathbf{v}) + \frac{2L}{K}\|\mathbf{w}^0 - \mathbf{v}\|^2 \leq F(\mathbf{v}) + \frac{4L}{K}\|\mathbf{v}\|^2 + \frac{4L}{K}\|\mathbf{w}^0\|^2. \tag{4}$$

We want to pick a proper $\mathbf{v}$ so that $g(\mathbf{v}) \triangleq F(\mathbf{v}) + \frac{4L}{K}\|\mathbf{v}\|^2$ is small.

Since this is the linearly separable case, i.e., there exists a sequence $\mathbf{v}^K$ such that $F(\mathbf{v}^K) \to 0$; but meanwhile, the norm $\|v^K\| \to \infty$, thus there is a trade-off between $F(\mathbf{v})$ and $\|\mathbf{v}\|^2$. How to pick a proper $\mathbf{v}$?

We consider the separating vector $\mathbf{w}^*$: suppose $\mathbf{w}^*$ satisfies $\|\mathbf{w}^*\| = 1$ and $y_i \mathbf{x}_i^\top \mathbf{w}^* \geq \gamma > 0, \forall\, i$. We let $\mathbf{v} = \mathbf{w}^* \ln(c)/\gamma$ where $c > 1$, then $F(\mathbf{v}) = \sum_i \ln(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w}^* \ln c/\gamma)) \leq \sum_i \ln(1 + \exp(-\ln c\ )) = n \ln(1 + \frac{1}{c}) \leq \frac{n}{c}$.
Thus

$$g(\mathbf{v}) \leq \frac{n}{c} + \frac{4L}{K}\frac{\ln(c)^2}{\gamma^2}$$

Pick $c = Kn\gamma^2$, then

$$g(\mathbf{v}) \leq \frac{1 + 4L\ln(Kn\gamma^2)^2}{K\gamma^2}.$$

Combining with (4), then we are done.

Nonlinearly Separable Case

Assumption 2 (non-linearly separable): There does not exist any $\mathbf{w}^* \in \mathbb{R}^d$ such that $y_i \mathbf{x}_i^\top \mathbf{w}^* \geq 0, \forall\, i = 1, 2, \ldots, n$.

Claim 1: Under Assumption 2, $F(\mathbf{w})$ is coercive. (Thus there exists at least one global minimum.)

Claim 2: $F(\mathbf{w})$ is NOT strongly convex.

Convergence result (informal): GD with a proper stepsize converges to global optima at a linear rate.

Sketch of analysis:

- Since $F(\mathbf{w})$ is coercive, any level set is bounded.
- GD will stay in a level set, thus stay in a bounded region.
- $F(\mathbf{w})$ is a strongly convex function in this bounded region; thus GD converges to global-min at a linear rate.

Proofs of Claims (reading)

Proof of Claim 1 (coercive): Consider the function $\phi(\mathbf{u}) = \min_{1\leq j\leq n} y_j \mathbf{x}_j^\top \mathbf{u}$ on the surface $S^d \triangleq \{\mathbf{u} \in \mathbb{R}^d : \|\mathbf{u}\| = 1\}$. According to Assumption 2, for any $\mathbf{u}$, there exists at least one j such that $y_j \mathbf{x}^\top \mathbf{u} < 0$, thus $\phi(\mathbf{u}) < 0, \forall \mathbf{u} \in S^d$. Since $g(\mathbf{u})$ a continuous function defined on a compact set, thus it has a global maximal value, denoted as $-\tau^*$. In other words, for any $\mathbf{u} \in S^d$, there exists at least one j such that $y_j \mathbf{x}_j^\top \mathbf{u} \leq -\tau^*$ Consider an arbitrary sequence $\mathbf{w}(k), k = 1, 2, \ldots$ where $\lim_{k\to\infty} \|\mathbf{w}(k)\| = \infty$. Denote $\beta_k = \|\mathbf{w}(k)\|$ and $\mathbf{v}(k) = \mathbf{w}(k)/\|\mathbf{w}(k)\| \in S^d$, then $\lim_{k\to\infty} \beta_k = \infty$, and $\min_{1\leq j\leq n} y_j \mathbf{x}_j^\top \mathbf{v}(k) \leq -\tau^*$. We denote $j(k) = \mathrm{argmin}_j\, y_j \mathbf{x}_j^\top \mathbf{v}(k)$, then $y_{j(k)} \mathbf{x}_{j(k)}^\top \mathbf{v}(k) \leq -\tau^*$. Therefore, $F(\mathbf{w}(k)) = \sum_{i=1}^n \log(1 + \exp(-y_i \mathbf{x}_i^\top \mathbf{w})) \geq \log(1 + \exp(-\beta_k y_{j(k)} \mathbf{x}_{j(k)}^\top \mathbf{v}(k))) \geq \log(1 + \exp(\beta_k \tau^*))$. Since $\lim_{k\to\infty} \log(1 + \exp(\beta_k \tau^*)) = \infty$, we have $\lim_{k\to\infty} F(\mathbf{w}(k)) = \infty$.

Proof of Claim 2 (not strongly convex):
$\nabla^2 F(\mathbf{w}) = \sum_{i=1}^n \frac{e^{-y_i w^\top x_i}}{(1+e^{-y_i w^\top x_i})^2} x_i x_i^\top$.
Note that $\lim_{t\to\infty} \frac{e^t}{(1+e^t)^2} = 0$, and $\lim_{t\to-\infty} \frac{e^t}{(1+e^t)^2} = 0$.
There exists at least one vector $\mathbf{u}$ such that $\mathbf{u}^\top \mathbf{x}_i \neq 0, \forall i$. Consider a sequence $\mathbf{v}(k) = k\mathbf{u}, k = 1, 2, 3, \ldots$. Then $\lim_{k\to\infty} \mathbf{v}(k)^\top \mathbf{x}_i \in \{\infty, -\infty\}$. Then $\lim_{k\to\infty} \frac{e^{-y_i \mathbf{v}(k)^\top x_i}}{(1+e^{-y_i \mathbf{v}(k)^\top x_i})^2} = 0, \forall i$. Thus $\lim_{k\to\infty} \nabla^2 f(\mathbf{v}(k)) = 0$. Thus there cannot exist any positive constant μ such that $\lambda_{\min}(\nabla^2 F(\mathbf{w})) \geq \mu, \forall \mathbf{w}$. Thus F is not strongly convex.

Outline

Warmup: 1-dim Linear Regression

1-dim logistic regression

- Logistic Regression Introduction
- First Data Setting, Stopping Criteria
- Second Data Setting

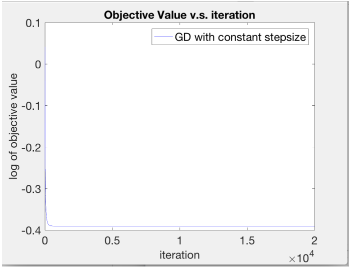
Formal Results for Logistic Regression

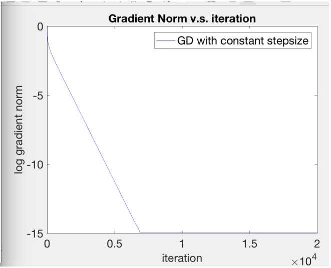
Experiments of Higher-Dim LR

Higher-Dim LR: Non-Separable Case

Data Setting 1: $n = 10, d = 5$.

- Random Ganssian data $x_1, \ldots, x_{10} \in \mathbb{R}^5$.
- Random labels y_i with prob. 1/2.





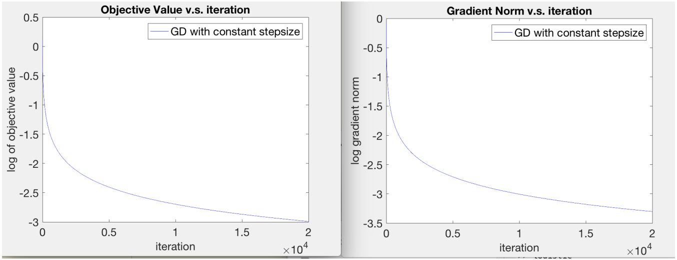
Explain?

- In this case, objective is strongly convex

Higher-Dim LR: Separable Case

Data Setting 2: $N = 10, d = 5$.

- Random Ganssian data $x_1, \ldots, x_{10} \in \mathbb{R}^5$.
- True separator $w^* = (1, 1, 1, 1, 1)$
- $y_i = \text{sign}(w' x_i), \forall i$.



Explain?

- In this case, objective is non-strongly convex

Applying Theory to Explain More Details

If you want, we could explain many more details of the convergence behavior of LR.

e.g. in Data Setting 1, what determines the rate of convergence?

e.g. in Data Setting 2, it converges in $O(1/K)$ rate, what is the constant?

In higher-dimensional case, like $n, d = 1000$ or higher, the behavior may be more complicated.

- Again, if you are patient enough, you can explain the behavior

This confidence should NOT be extended to complicated non-convex problem (e.g. deep learning): we could NOT explain many behaviors in practice.

Side knowledge: High-Dim Case?

Separable case causes convergence issues – this was known for a long time.

- Allison 2008, “Convergence Failures in Logistic Regression.”
- “A frequent problem in estimating logistic regression models is a failure of the ... algorithm to converge.”
  - Here “failure of converge” means “failure of iterates to converge”
  - We know that the function values will converge
- “Two possible data patterns: complete separation, or quasi-complete separation”. The second is very common.

Resolve Issues in Separable Case

We mention two possible solutions.

Solution 1: Add regularizer, e.g., $\lambda\|\mathbf{w}\|^2$.

- Claim: If f is convex, then $f(\mathbf{w}) + \lambda\|\mathbf{w}\|^2$ is strongly convex.
- Issues: how to tune λ? Distort original solutions? Still slow....

Solution 2: Add constraints on w, or normalize w during algorithms (beyond the current class).

Some Practical Lessons

Even if the Hessian at every iterate is well-conditioned, the algorithm might converge slowly

- may need to compare eigenvalues across iterates

Check the norm of the weights: if it goes very large, then it may be the reason for slow convergence (for problems similar to logistic regression)

Blog post and reference

Reference (on sublinear convergence): Sec 7.1.2 and Sec. 10.2 of Matus's lecture notes `https://mjt.cs.illinois.edu/dlt/`

Related blog:
A former student of UIUC wrote a blog post:

Gradient descent in logistic regression: `https://roth.rbind.io/post/gradient-descent-in-logistic-regression/`

A bit funny story: I want to search who has formally analyzed optimization of LR, and finds this blog that seems great.

But when I check reference, it refers to my own teaching slides...

Summary of This Lecture

- **1-dim quadratic** problem: stepsize and different phases
  - Verify the theoretical threshold $2/L$
  - Verify the optimal stepsize $2/(L+\mu)$

- Stopping criterion, and plotted metric
  - Plot $f(\mathbf{x}_r)$; common in practice
  - Plot $\|\nabla f(\mathbf{x}_r)\|$ or $\log(\|\nabla f(\mathbf{x}_r)\|)$; better metric in terms of detecting convergence

- **Optimization analysis of LR** (logistic regression)
  - Sublinear rate $O(\ln(K)/K)$ for linearly separable case
  - Linear rate for non-linearly seprable case (no formal proof presented)

- Lessons from 1-dim logistic regression experiments:
  - sub-linear rate for convex (non-strongly convex) problem really happens in practice
  - condition number: maximal any-Hessian-eigenvalue over minimal any-Hessian-eigenvalue. Can be arbitrarily large even for 1D strictly convex function
  - Separable v.s. non-separable: sublinear v.s. linear