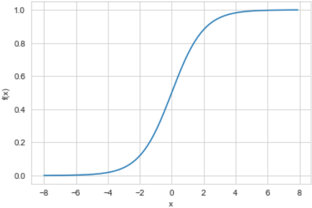
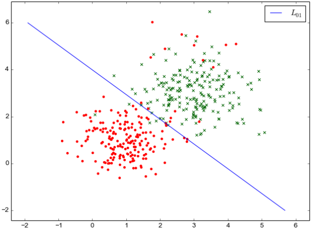
Worst Case Datasets for Solving Binary Logistic Regression via Deterministic First-Order Methods

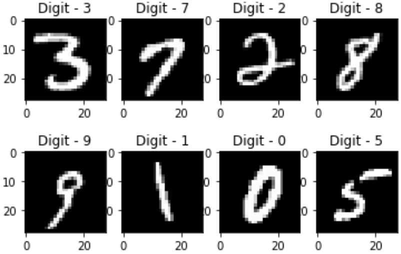
Trevor Squires[1]

Thesis Defense, July 21, 2021

[1]Supported by the National Science Foundation through grant DMS-1913006.







Binomial Logistic Regression

Introduction

Binary Logistic Regression

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

- a_i^T represent rows of data matrix $A \in \mathbb{R}^{N\times n}$
- $b_{(i)}$ are the entries of the response vector $b \in \{-1,1\}^N$
- Assumes $P(b_{(i)} = 1 \mid a_i^T; x, y) = \frac{1}{1+\exp(-a_i^T x + y)}$
- Model formulated by maximum likelihood estimation
- $n >> 1$

Introduction

Can we work with something cleaner? Define

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

for any $u \in \mathbb{R}^k$.

$$
\phi^*_{A,b} := \min_{x \in \mathbb{R}^n, y \in \mathbb{R}} \phi_{A,b}(x, y) := h(Ax + y1) - b^T(Ax + y1)
$$

Our goal: compute an ε-solution $(\hat{x}, \hat{y})$ such that $\phi_{A,b}(\hat{x}, \hat{y}) - \phi^*_{A,b} < \varepsilon$ as quickly as possible.

Introduction

Goal

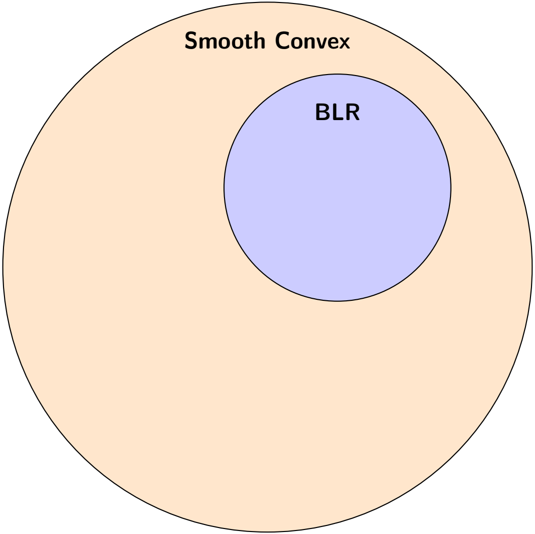
Compute an ε-solution to

$$\phi^*_{A,b} := \min_{x \in \mathbb{R}^n, y \in \mathbb{R}} \phi_{A,b}(x,y) := h(Ax + y1) - b^T(Ax + y1)$$

as quickly as possible.

- $\phi_{A,b}(x,y)$ has the following properties:
  - $\phi_{A,b}(x,y)$ is convex
  - $\nabla \phi_{A,b}(x,y)$ is Lipschitz continuous
- Solving for $\phi^*_{A,b}$ is smooth, convex, and unconstrained optimization
- Can relax to just solving smooth, convex problems

Smooth Convex Optimization



Smooth Convex Optimization

Goal

Compute an ε-solution to
$$f^* := \min_{x \in \mathbb{R}^n} f(x)$$
as quickly as possible. Here, f is convex differentiable and has L-Lipschitz continuous gradient, i.e. $||\nabla f(x) - \nabla f(y)|| \leq L\,||x - y||\,, \forall x, y \in \mathbb{R}^n$.

- Large class of problems
- Examples include
  - Regularized Linear Least Squares
$$f(x) = \frac{1}{2}\,||Ax - b||^2 + \lambda\,||x||^2$$
  - Quadratic Programming
$$f(x) = \frac{1}{2} x^T A x - b^T x, A \succeq 0$$
- What is a first order method?
  - any method $\mathcal{M}$ such that $\mathcal{M}$ 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 $x \in \mathbb{R}^n$

Algorithm 1 Nesterov's accelerated gradient descent (NAGD)

Select parameters $\gamma_k \in (0,1]^N, \eta_k$. Choose $x_0 \in \mathbb{R}^n$. Set $y_0 = x_0$.

for $k = 1, \ldots, N$ **do**

$$\begin{aligned} z_k =& (1-\gamma_k)y_{k-1} + \gamma_k x_{k-1} \\ x_k =& \operatorname*{argmin}_{x \in \mathbb{R}^n} \langle \nabla f(z_k), x \rangle + \frac{\eta_k}{2} ||x_{k-1} - x||_2^2 \\ y_k =& (1-\gamma_k)y_{k-1} + \gamma_k x_k \end{aligned}$$

end for

Output y_N.

- Depends on parameters γ_k, η_k.
- Different parameter settings = different performance

Smooth Convex Optimization

Goal

Compute an ε-solution to

$$f^* := \min_{x \in \mathbb{R}^n} f(x)$$

as quickly as possible. Here, f is convex differentiable and has L-Lipschitz continuous gradient, i.e. $||\nabla f(x) - \nabla f(y)|| \leq L\,||x - y||\,, \forall x, y \in \mathbb{R}^n$.

If we set $\gamma_k \equiv 1$ and $\eta_k \equiv L$ in NAGD, then

- $x_k = (x_{k-1} - \frac{1}{L}\nabla f(x_{k-1}))$
- NAGD reduces to gradient descent (GD)
- $f(\tilde{y}_N) - f(x^*) \leq \frac{L||x^* - x_0||^2}{N+1}$ where $\tilde{y}_N = \sum_{k=0}^{N} y_k/(N+1)$
- Computes an ε-solution in $\mathcal{O}(1/\varepsilon)$ iterations

GD provides an upper complexity bound of $\mathcal{O}(1/\varepsilon)$ for smooth convex optimization. Is this "as quickly as possible?"

Smooth Convex Optimization

Goal

Compute an ε-solution to
$$f^* := \min_{x \in \mathbb{R}^n} f(x)$$
as quickly as possible. Here, f is convex differentiable and has L-Lipschitz continuous gradient, i.e. $||\nabla f(x) - \nabla f(y)|| \leq L \, ||x - y|| \, , \forall x, y \in \mathbb{R}^n$.

If we set $\gamma_k = \frac{2}{k+1}$ and $\eta_k = \frac{2L}{k}$ in NAGD, then

- $f(y_N) - f(x^*) \leq \frac{4L}{N(N+1)} \, ||x^* - x_0||^2$
- Computes an ε-solution in $\mathcal{O}(1/\sqrt{\varepsilon})$ iterations
- Asymptotically better than gradient descent
- Called Optimal Gradient Descent (OGD)

OGD provides an upper complexity bound of $\mathcal{O}(1/\sqrt{\varepsilon})$ for smooth convex optimization. Is this "as quickly as possible?"

Lower Complexity Bound

Goal

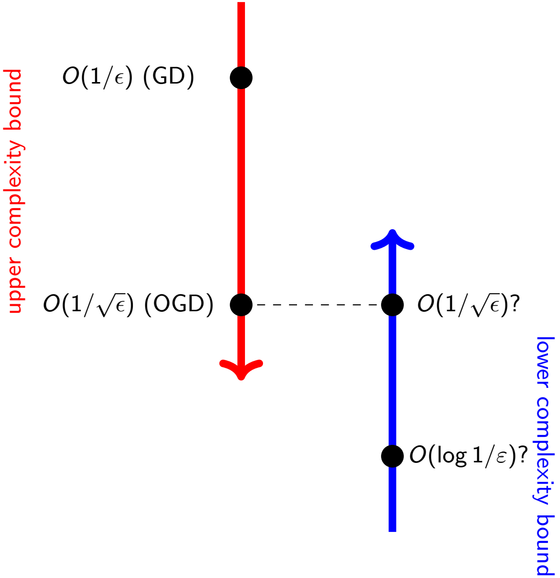
Compute an ε-solution to

$$f^* := \min_{x \in \mathbb{R}^n} f(x)$$

as quickly as possible.

- What does "as quickly as possible" mean?
- How can we evaluate the worst-case performance of an algorithm?
- Search for some "difficult" problem instance such that said algorithm struggles to solve it.
- A worst case problem instance for a class of algorithms provides a lower complexity bound.

Complexity Bounds



Algorithm Class Comparison

Goal

Compute an ε-solution to

$$f^* := \min_{x \in \mathbb{R}^n} f(x)$$

as quickly as possible.

- Why exactly do we consider iterative first order method?
- Consider a simple problem class: quadratic programming

$$\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T A x - b^T x, A \succeq 0$$

  - second order methods (Newton's) require 1 iteration of $\mathcal{O}(n^3)$ (requires linear system solve) flops
  - first order methods require t iterations of $\mathcal{O}(n^2)$ flops
  - If $t \leq n$, i.e. when n is large, first order seems best

Smooth Convex Optimization

Goal

Compute an ε-solution to

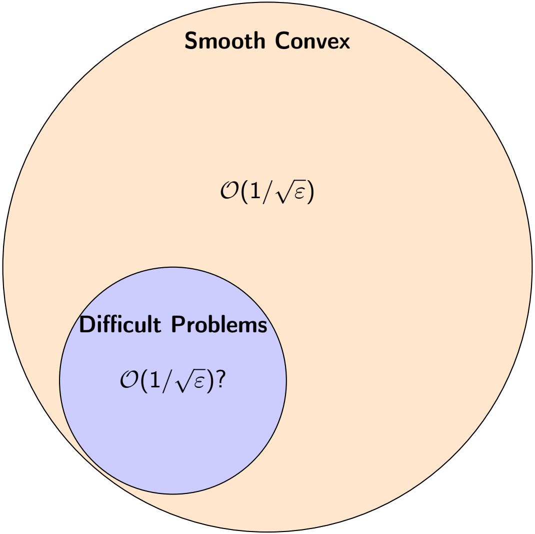
$$f^* := \min_{x \in \mathbb{R}^n} f(x)$$

as quickly as possible. Here, f is convex differentiable and has L-Lipschitz continuous gradient, i.e. $||\nabla f(x) - \nabla f(y)|| \leq L\, ||x - y||\,, \forall x, y \in \mathbb{R}^n$.

Let's review

- Binary logistic regression is in the class of smooth convex optimization problems
- Optimal gradient descent solves smooth convex optimization problems in $\mathcal{O}(1/\sqrt{\varepsilon})$ iterations
- We hope to find a problem instance such that no first order method can solve it faster than $\mathcal{O}(1/\sqrt{\varepsilon})$

Lower Complexity Bound Goal



Complexity Bounds

In [1], Nemirovski showed that the lower complexity bound of solving

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

via first order deterministic methods was $\mathcal{O}(1/\sqrt{\varepsilon})$, i.e. OGD is indeed optimal.

Key ideas from Nemirovski:

- Construct a worst-case instance of f such that any first order method $\mathcal{M}$ struggles to solve it.
- Find an "equivalent" function g such that all iterates x_t generated by $\mathcal{M}$ applied to g lie in a particular subspace.
- Show that the error at step t of $\mathcal{M}$ applied to g is at least as large as the proposed lower complexity bound.

Nemivroski+Nesterov Proof Sketch

Key Idea

Construct a worst-case instance of f such that any first order method $\mathcal{M}$ struggles to solve it.

- $A_{4k+3} = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}$
- $A = \frac{L}{4}\begin{pmatrix} A_{4k+3} & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{n\times n}$, $b = \frac{L}{4}e_1$
- $\min\limits_{x\in\mathcal{K}_{2k+1}(A,b)} Q_{A,b}(x) - \min\limits_{x\in\mathbb{R}^n} Q_{A,b}(x) \geq \frac{3L\left|\left|x^*\right|\right|^2}{128(k+1)^2}$
- Here, $\mathcal{K}_r(A,b) = \text{span}\{b, Ab, \ldots, A^{r-1}b\}$

If each iterate $x_t \in \mathcal{K}_{2k+1}(A,b)$, we are done!

Nemirovski+Nesterov Proof Sketch

Key Idea

Find an "equivalent" function g such that all iterates x_t generated by $\mathcal{M}$ applied to g lie in a particular subspace.

- If $x_t \notin \mathcal{K}_{2k+1}(A, b)$, we can rotate the problem, i.e. find $g(x) := f(Ux)$, such that
  - $x_t \in U^T \mathcal{K}_{2k+1}(A, b)$ for some orthogonal matrix U satisfying $Ub = b$
  - $\min_{x \in U^T \mathcal{K}_r(A,b)} Q_{U^T AU, b}(x) - \min_{x \in \mathbb{R}^n} Q_{U^T AU, b}(x) = \min_{x \in \mathcal{K}_r(A,b)} Q_{A,b}(x) - \min_{x \in \mathbb{R}^n} Q_{A,b}(x)$
- If g and f have the same first order information at the oracle query points, then $\mathcal{M}$ "cannot differentiate" between the two
- Utilizes an important lemma

Lemma

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

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

Nemirovski+Nesterov Proof Sketch

Key Idea

Show that the error at step t of $\mathcal{M}$ applied to the rotated objective function is at least as large as the proposed lower complexity bound.

Theorem

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

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

We conclude OGD is optimal for smooth convex optimization

Nemirovski vs. Nesterov Technique

Nemirovski:

- Constructed A by characterizing its spectrum
  - WLOG may assume A is diagonal since for $A = V^T \Lambda V$,

$$\min_{x\in\mathbb{R}^n} \frac{1}{2}x^T Ax - b^T x = \min_{x\in\mathbb{R}^n} \frac{1}{2}x^T V^T \Lambda V x - b^T V^T V x = \min_{y\in\mathbb{R}^n} \frac{1}{2} y^T \Lambda y - b^* y$$

- Enforced iterates in Krylov subspace using rotation/orthogonal invariance trick
- Pros: Rigorous, general
- Cons: Hard to follow, diagonalization may not hold in other setting

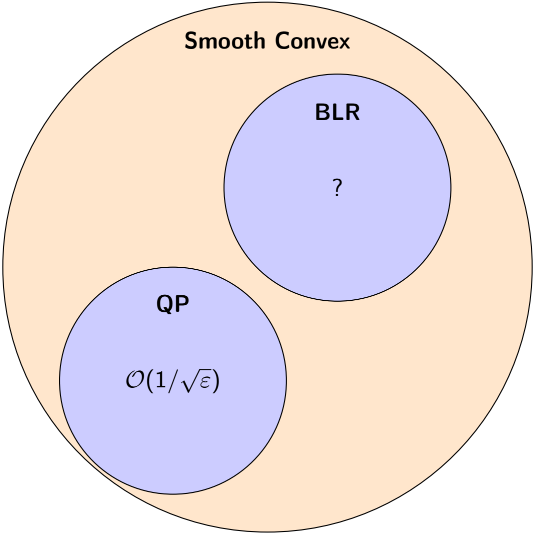
Nesterov:

- Constructed A using tridiagonal form
- Enforce iterates in Krylov subspace using linear span assumption (shown in [2])
- Pros: Easy to follow
- Cons: Requires assumption

Lower Complexity Extensions

- 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 $\mathcal{O}(1/\varepsilon^2)$ [1, 2]
  - when f is convex, nonsmooth with bilinear saddle point structure, the lower complexity bound is $\mathcal{O}(1/\varepsilon)$ [3]
  - when f is strongly convex, smooth the lower complexity bound is $\mathcal{O}(\log(1/\varepsilon))$ [2, 4]
- What about binary logistic regression?
  - can we do better than smooth convex optimization?
  - can we adapt Nemirovski/Nesterov's idea to binary logistic regression?

Lower Complexity Bound Summary



BLR Lower Complexity Bound

- Extend this result to binary logistic regression problems
- Construct a worst-case dataset for solving binary logistic regression that requires $\mathcal{O}(1/\sqrt{\varepsilon})$ first order oracle calls
- These worst-case constructions will satisfy $y^* = 0$. Consequently, it suffices to solve the logistic model with homogeneous linear predictor

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

and corresponding problem

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

- We assume that
  - (initially) the iterates of a deterministic first order method $\mathcal{M}$ satisfy $x_t \in \text{span}\{\nabla f(x_0), \ldots, \nabla f(x_{t-1})\}$
  - $x_0 = 0$
  - x_t's are inquiry points and approximate solutions

BLR Lower Complexity Bound

Binary Logistic Regression

$$l^*_{A,b} := \min_{x \in \mathbb{R}^n} l_{A,b}(x) := h(Ax) - b^T Ax$$

- Given any k, let $W_k := \begin{pmatrix} & & & -1 & 1 \\ & & -1 & 1 & \\ & \cdot^{\cdot^{\cdot}} & \cdot^{\cdot^{\cdot}} & & \\ -1 & 1 & & & \\ 1 & & & & \end{pmatrix} \in \mathbb{R}^{k \times k}$, $A_k := \begin{pmatrix} 2\sigma W_k \\ -2\zeta W_k \\ -2\sigma W_k \\ 2\zeta W_k \end{pmatrix} \in \mathbb{R}^{4k \times k}$, $b_k = \begin{pmatrix} 1_{2k} \\ -1_{2k} \end{pmatrix} \in \mathbb{R}^{4k}$ and $\sigma > \zeta > 0$.
- Define $f_k(x) := h(A_k x) - b_k^T(A_k x)$ and $\phi_k(x, y) := h(A_k x_+ y 1_k) - b_k^T(A_k x + y 1_k)$
- $x^* = \operatorname*{argmin}_{x \in \mathbb{R}^k} f_k = c(1, 2, \ldots, k)^T$
- $f^*_k = f_k(x^*) = 8k \log 2 + 4k \left(\log\cosh(\sigma c) + \log\cosh(\zeta c) - (\sigma - \zeta)c\right)$

Objective Functions

$$W_k := \begin{pmatrix} & & & -1 & 1 \\ & & -1 & 1 & \\ & \iddots & \iddots & & \\ -1 & 1 & & & \\ 1 & & & & \end{pmatrix}, A_k := \begin{pmatrix} 2\sigma W_k \\ -2\zeta W_k \\ -2\sigma W_k \\ 2\zeta W_k \end{pmatrix}, b_k = \begin{pmatrix} \mathbf{1}_{2k} \\ -\mathbf{1}_{2k} \end{pmatrix}$$

- Define for any positive integers t and k

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

and

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

- $W_k 1_k = e_{k,k}, A_k^T b_k \in \mathcal{X}_{k,k}$
- For $x = \begin{pmatrix} 0_{k-t} \\ u \end{pmatrix} \in \mathcal{X}_{t,k},\ W_k x = \begin{pmatrix} W_t u \\ 0_{k-t} \end{pmatrix}$, and $A_k x, \nabla h(A_k x) \in \mathcal{Y}_{t,k}$
- For $y = \begin{pmatrix} v \\ 0_{k-t} \end{pmatrix} \in \mathcal{X}^C_{k-t,k},\ W_k^T v = \begin{pmatrix} 0_{k-t-1} \\ -v_{(t)} \\ W_t v \end{pmatrix} \in \mathcal{X}_{t+1,k}$
- $A_k^T \nabla h(A_k x) = 4\sigma W_k^T \begin{pmatrix} \tanh(\sigma W_t u) \\ 0_{k-t} \end{pmatrix} + 4\zeta W_k^T \begin{pmatrix} \tanh(\zeta W_t u) \\ 0_{k-t} \end{pmatrix} \in \mathcal{X}_{t+1,k}$

Properties of W_k and A_k

Linear Span Assumption

When $\mathcal{M}$ is applied to solve f_k, the iterates x_t generated by $\mathcal{M}$ satisfy

$$x_t \in \text{span}\{\nabla f_k(x_0), \ldots, \nabla f_k(x_{t-1})\}$$

- Recall: $A_k^T b_k \in \mathcal{X}_{k,k}$
- Recall: $A_k^T \nabla h(A_k x) \in \mathcal{X}_{t+1,k}$
- $\nabla f_k(x_t) = A_k^T \nabla h(A_k x) - A_k^T b_k \in \mathcal{X}_{t+1,k}$
- The linear span assumption gives $x_t \in \mathcal{X}_{t,k}$ "for free"
- Can compute $\min_{x \in \mathcal{X}_{t,k}} f_k(x) - f_k^* = 8(k-t)\log 2 + f_t^* - f_k^*$

BLR Lower Complexity Bound

Objective Function

$$l_{A,b}^* := \min_{x \in \mathbb{R}^n} l_{A,b}(x) := \min_{x \in \mathbb{R}^n} h(Ax) - b^T Ax$$

Theorem

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

$$l_{A,b}(x_T) - l_{A,b}(x^*) \geq \frac{3 \, ||A||^2 \, ||x_0 - x^*||^2}{32(2T+1)(4T+1)}$$

and

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

BLR Lower Complexity Bound

Key ideas from Nemirovski:

- Construct a worst-case instance of f such that any first order method $\mathcal{M}$ struggles to solve it. Done via A_k, W_k, and b_k similar to Nesterov
- Find an "equivalent" function g such that it shares the first order information of f and all iterates x_t generated by $\mathcal{M}$ applied to g lie in a particular subspace. Done using f_k via linear span assumption
- Show that the error at step t of $\mathcal{M}$ applied to g is at least as large as the proposed lower complexity bound. Done in the same way as Nemirovski

Do we need the linear span assumption, i.e. can we find a related function g similar to Nemirovski?

Lemma

For A_k, b_k specified previously, 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

- $U_t A_k^T b_k = A_k^T b_k$
- *When $\mathcal{M}$ is applied to solve $l_{A_k U_t, b_k}$, the iterates $x_0, \ldots, x_t$ satisfy*

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

- Idea: use successive instances of the rotation lemma to find matrices that fix all previous iterates **and** places the next iterate in a larger subspace
- Show that a first order algorithm "can not tell a difference" of the original problem and the rotated problem, i.e. they have the same first order information

BLR Lower Complexity Bound

Objective Function

$$l_{A,b}^* := \min_{x \in \mathbb{R}^n} l_{A,b}(x) := \min_{x \in \mathbb{R}^n} h(Ax) - b^T Ax$$

Theorem

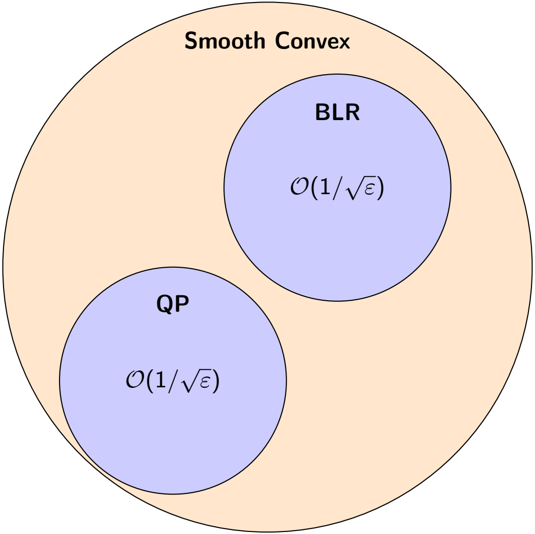
(Presented in [5]) For any first order method $\mathcal{M}$ and fixed iteration number T with corresponding constants $N = 10T + 8, n = 4T + 2$, there always exists data matrix $A \in \mathbb{R}^{N \times n}$ and response vector $b \in \mathbb{R}^N$ such that when $\mathcal{M}$ is applied to solve $l_{A,b}$, the T-th iterate satisfies

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

and

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

Lower Complexity Bound Summary



Concluding Remarks

- Conditions
  - First order oracle assumption
  - Large dimensionality assumption
- Unconstrained quadratic optimization of the form

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

  has a lower bound complexity of $\mathcal{O}(1/\sqrt{\varepsilon})$
  - OGD is optimal for smooth convex optimization
  - CG is optimal for unconstrained quadratic optimization
- (Homogeneous) Binary logistic regression of the form

$$\min_{x \in \mathbb{R}^n} h(Ax) - b^T (Ax)$$

  has a lower bound complexity of $\mathcal{O}(1/\sqrt{\varepsilon})$
  - OGD is optimal for homogeneous binary logistic regression
  - OGD is optimal for inhomogeneous binary logistic regression

References

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

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

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.

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.

Yuyuan Ouyang and Trevor Squires.
Some worst-case datasets of deterministic first-order methods for solving binary logistic regression.
Inverse Problems and Imaging, 2019.