CS6301: Optimization in Machine Learning

Lecture 6 - 8: Gradient Descent and Family

Rishabh Iyer

Department of Computer Science
University of Texas, Dallas
https://sites.google.com/view/cs-6301-optml/home

February 3rd, 2020

Project and Assignment

- Project Deadline 1: Finalize on your Project Topics and partners: **February 15th 2020**
- Projects can be done in Groups with 1-3 students per group
- Project Deadline 2: Mid Term Review of the Project: **March 15th 2020**
- Final Project Report Deadline: **April 20th 2020**
- Last 3-5 Lectures of this class will be the course project presentations. Around 10 mins per project.
- **Updated Assignment Posted on eLearning. Due Date now is 5th February**

Outline

- Recap from Previous Lecture
- Recap on Local and Global Extrema
- Lipschitz Continuity, Strong Convexity and Lipschitz Smoothness
- Gradient Descent and Analysis: Continuous, Smooth and Strong Convex (and their combinations)
- Accelerated Gradient Descent and Lower Bounds

Recap: Convex Functions

- A Function $f : \mathbb{R}^d \to \mathbb{R}$ is convex if:
  - $dom(f)$ is a convex set
  - for all $x, y \in dom(f)$ and $\lambda : 0 < \lambda < 1$, we have: $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$
- Geometrically, the line segment between $(x, f(x))$ and $(y, f(y))$ lies above the graph of f.

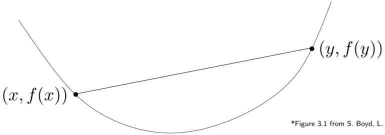


*Figure 3.1 from S. Boyd, L. Vandenberghe

- f is strictly convex if for all $x, y \in dom(f)$ and $\lambda : 0 < \lambda < 1$, we have: $f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$

Recap: Strongly Convex Functions

- A Function $f : \mathbb{R}^d \to \mathbb{R}$ is strongly convex if there exists a $\mu > 0$ such that the function $g(x) = f(x) - \mu/2\|x\|^2$ is convex
- The parameter μ is the strong convexity parameter
- Geometrically, strong convexity means that there exists a quadratic lower bound on the growth of the function.
- Its easy to see that Strong Convexity implies Strict Convexity!
- Strong Convexity Doesn't imply the function is differentiable!
- If a function f is strongly convex and g is convex (not necessarily strongly convex), $f + g$ is strongly convex.
- $\|x\|^2$ is strongly convex!
- Hence for any convex function f, the function $f(x) + \lambda/2\|x\|^2$ is strongly convex!

Recap: Which of the Following Loss Functions are Convex?

- L1/L2 Reg Logistic Regression: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i)) + \lambda\|\theta\|$
- L1/L2 Reg SVMs: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\} + \lambda\|\theta\|$
- L1/L2 Reg Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\} + \sum_{i=1}^{c} \lambda \sum_{j=1}^{m} \|\theta_j\|$
- L1/L2 Reg Least Squares (Lasso): $L(\theta) = \sum_{i=1}^{n}(\theta^T x_i - y_i)^2 + \lambda\|\theta\|$
- Matrix Completion: $L(X) = \sum_{i=1}^{n} ||y_i - A_i(X)||_2^2 + ||X||_*$
- Soft-Max Contextual Bandits: $L(\theta) = \sum_{i=1}^{n} \frac{r_i}{p_i} \frac{\exp(\theta^T x_i^{a_i})}{\sum_{j=1}^{k} \exp(\theta^T x_i^j)} + \lambda\|\theta\|$

Recap: First-Order Convexity Conditions

Theorem

1. *For differentiable $f : \mathcal{D} \to \Re$ and convex set $\mathcal{D}$, f is convex* **iff**, *for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

2. *f is strictly convex* **iff**, *for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, with $\mathbf{x} \neq \mathbf{y}$,*

$$f(\mathbf{y}) > f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x})$$

3. *f is strongly convex* **iff**, *for any $\mathbf{x}, \mathbf{y} \in \mathcal{D}$, and for some constant $c > 0$,*

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla^T f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{1}{2}c||\mathbf{y} - \mathbf{x}||^2$$

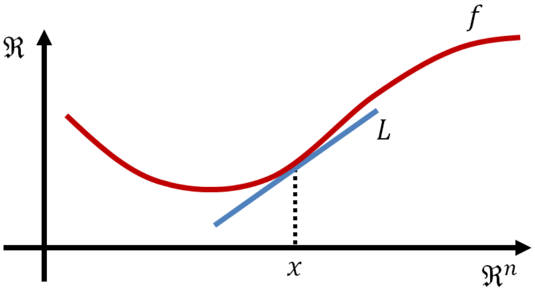
Recap: Second Order Conditions of Convexity

- Recall the Hessian of a continuous function:

$$\nabla^2 f(w) = \begin{pmatrix} \frac{\partial^2 f}{\partial w_1^2} & \frac{\partial^2 f}{\partial w_1 \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_1 \partial w_n} \\ \frac{\partial^2 f}{\partial w_2 \partial w_1} & \frac{\partial^2 f}{\partial w_2^2} & \cdots & \frac{\partial^2 f}{\partial w_2 \partial w_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial w_n \partial w_1} & \frac{\partial^2 f}{\partial w_n \partial w_2} & \cdots & \frac{\partial^2 f}{\partial w_n^2} \end{pmatrix}$$

- f is convex if and only if, a) $dom(f)$ is convex, and for all $x \in dom(f)$, $\nabla^2 f(x) \geq 0$ (i.e. $\nabla^2 f(x)$ is positive semi-definite).

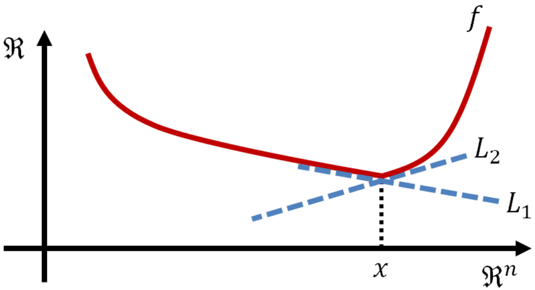
Recap: (Sub)Gradients and Convexity



To say that a function $f : \Re^n \mapsto \Re$ is differentiable at $\mathbf{x}$ is to say that there is a single unique linear tangent that under estimates the function:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \triangledown f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{x}, \mathbf{y}$$

Recap: (Sub)Gradients and Convexity

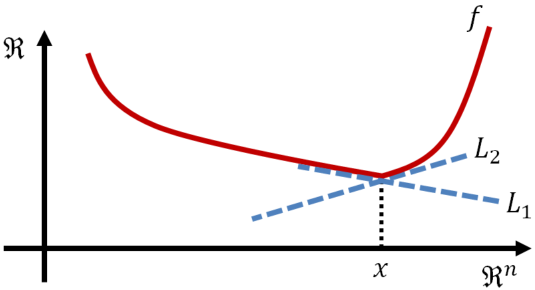


In this figure we see the function f at $\mathbf{x}$ has many possible linear tangents that may fit appropriately. Then a **subgradient** is any $\mathbf{h} \in \Re^n$ (same dimension as x) such that:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{y}$$

Thus, intuitively, if a function is differentiable at a point $\mathbf{x}$ then

Recap: (Sub)Gradients and Convexity



In this figure we see the function f at $\mathbf{x}$ has many possible linear tangents that may fit appropriately. Then a **subgradient** is any $\mathbf{h} \in \Re^n$ (same dimension as x) such that:

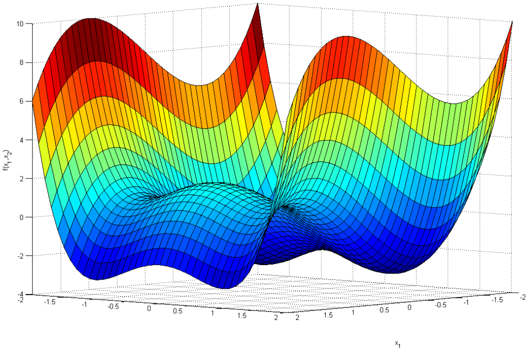
$$f(\mathbf{y}) \geq f(\mathbf{x}) + \mathbf{h}^T(\mathbf{y} - \mathbf{x}), \ \forall \mathbf{y}$$

Thus, intuitively, if a function is differentiable at a point $\mathbf{x}$ then it has a unique subgradient at that point ($\nabla f(\mathbf{x})$). Formal Proof?

Recap: Convexity and Continuity

- Let f be a convex function and suppose $dom(f)$ is open. Then f is continuous.
- How *wild* can non-differentiable convex functions be?
- While there are continuous functions which are nowhere differentiable, (see `https://en.wikipedia.org/wiki/Weierstrass_function`), convex functions cannot be pathological!
- Infact, a convex function is differentiable *almost* everywhere. In other words, the set of points where f is non-differentiable is of measure 0.
- However we cannot ignore the non-differentiability, since a) the global minima could easily be a point of non differentiability and b) with any optimization algorithms, you can stumble upon these "kinks".

Figure below shows the plot of $f(x_1, x_2) = 3x_1^2 - x_1^3 - 2x_2^2 + x_2^4$. As can be seen in the plot, the function has several local maxima and minima.



Recap: Local Minima

- If a function f is differentiable, and x is a local minima, then $\nabla f(x) = 0$.
- If f is not differentiable, then there counld be a local minima x with non-zero (sub)-gradient. Example: $f(x_1, x_2) = |x_1 - x_2|$. However, we can say that if x is a local minima, then $0 \in \partial f(x)$.
- Is the converse true? I.e. if x is s.t. $\nabla f(x) = 0$, then x is a local minima of f?
- No. For example, $f(x_1, x_2) = x_1^2 - x_2^2$. Such points are called saddle points!

Recap: Convexity and Global Minimum

Fundamental characteristics:

1. Any point of local minimum point is also a point of global minimum.
2. For any stricly convex function, the point corresponding to the gobal minimum is also unique.

Does Global Minima Always Exist?

- Does the global minimum always exist?
- Not necessarily even if f is bounded from below (e.g. $f(x) = e^x$)
- Weierstrass Theorem: Let f be a convex function and suppose there is a nonempty and bounded sublevel set $L_\alpha(f)$. Then f has a global minima.
- Since f is continuous, it attains a minimum over a closed and bounded (= compact) set $L_\alpha(f)$ at some x^*. Note that x^* is also a global minimum as firstly, $f(x^*) \leq f(x), \forall x \in L_\alpha(f)$. Next since, $f(x^*) \leq \alpha$, it follows that for any $x \notin L_\alpha(f)$, $f(x) > \alpha \geq f(x^*)$

Critical Points are Global Minima for Convex Functions

- Lemma: Suppose that f is convex and differentiable over an open domain $dom(f)$. Let $x \in dom(f)$. Then if $\nabla f(x) = 0$ (i.e. a critical point), then x is a global minima.
- Proof: Suppose $\nabla f(x) = 0$. Then from the first order characterization of convex functions, $\forall y \in dom(f), f(y) \geq f(x) + \nabla f(x)^T(y - x) \geq f(x)$. Hence x is a global minima.
- Note that this cannot be extended to non-differentiable convex functions since the global minima may not be a differentiable point (for example: $f(x) = ||x||_1$).

Convex Optimization Problem

- Formally, a convex optimization problem is an optimization problem of the form

$$\begin{aligned} &minimize \ f(\mathbf{w}) \\ &subject \ to \ c \ \in \ C \end{aligned}$$

  where f is a convex function, X is a convex set, and $\mathbf{w}$ is the optimization variable.
- if $X = dom(f)$, this becomes unconstrained optimization.
- A special case (f is a convex function, g_i are convex functions, and h_i are affine functions, and $\mathbf{x}$ is the vector of optimization variables):

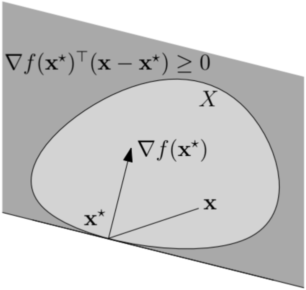
$$\begin{aligned} minimize \ f(\mathbf{w}) & \\ subject \ to \ g_i(\mathbf{w}) &\leq 0, \ i = 1, ..., m \\ h_i(\mathbf{w}) &= 0, i = 1, ..., p \end{aligned}$$

Optimality Conditions for Constrained Optimization

- Lemma: Suppose that f is convex and differentiable over an open domain $dom(f)$. Let $X \subseteq dom(f)$ be a convex set. A point x^* is a minimizer of f over X if and only if

$$\nabla f(x^*)^T(x - x^*) \geq 0, \forall x \in X$$

- Note that the Condition for Unconstrained minimization becomes a special case.
- Nice geometric interpretation:



Linear and Quadratic Programs

- Linear Program (LP) is a special case of a convex optimization problem:

$$\begin{aligned} &\textit{minimize } c^T x \\ &\textit{subject to } Ax \leq b \end{aligned}$$

- Another special case is Quadratic Programs (QP):

$$\begin{aligned} &\textit{minimize } 1/2x^T Qx \\ &\textit{subject to } Ax \leq b \end{aligned}$$

- The QP is a convex optimization problem only if Q is positive semi-definite,

Lipschitz Continuity

- A function f is Lipschitz continuous with Lipschitz constant L if

$$|f(x) - f(y)| \leq L||x - y||$$

Lipschitz Continuity

- A function f is Lipschitz continuous with Lipschitz constant L if

$$|f(x) - f(y)| \leq L||x - y||$$

- Basically Lipschitz continuity limits how fast a function changes.

Lipschitz Continuity

- A function f is Lipschitz continuous with Lipschitz constant L if

$$|f(x) - f(y)| \leq L||x - y||$$

- Basically Lipschitz continuity limits how fast a function changes.
- Lipschitz continuity implies that any $x \in dom(f)$ the (sub)gradient $h \in \partial f(x)$ satisfies $||h|| \leq L$.

Lipschitz Continuity

- A function f is Lipschitz continuous with Lipschitz constant L if

$$|f(x) - f(y)| \leq L||x - y||$$

- Basically Lipschitz continuity limits how fast a function changes.
- Lipschitz continuity implies that any $x \in dom(f)$ the (sub)gradient $h \in \partial f(x)$ satisfies $||h|| \leq L$.
- Properties of Lipschitz continuity:

Lipschitz Continuity

- A function f is Lipschitz continuous with Lipschitz constant L if

$$|f(x) - f(y)| \leq L||x - y||$$

- Basically Lipschitz continuity limits how fast a function changes.
- Lipschitz continuity implies that any $x \in dom(f)$ the (sub)gradient $h \in \partial f(x)$ satisfies $||h|| \leq L$.
- Properties of Lipschitz continuity:
  - If f_1 is L_1-Lipschitz continuous and f_2 is L_2-Lipschitz continuous, then $f_1 + f_2$ is $L_1 + L_2$ Lipschitz continuous

Lipschitz Continuity

- A function f is Lipschitz continuous with Lipschitz constant L if

$$|f(x) - f(y)| \leq L||x - y||$$

- Basically Lipschitz continuity limits how fast a function changes.
- Lipschitz continuity implies that any $x \in dom(f)$ the (sub)gradient $h \in \partial f(x)$ satisfies $||h|| \leq L$.
- Properties of Lipschitz continuity:
  - If f_1 is L_1-Lipschitz continuous and f_2 is L_2-Lipschitz continuous, then $f_1 + f_2$ is $L_1 + L_2$ Lipschitz continuous
  - Product of two Lipschitz continuous and bounded functions is also Lipschitz continuous.

Smooth Functions

- A function f is called smooth if its gradient is Lipschitz continuous.

Smooth Functions

- A function f is called smooth if its gradient is Lipschitz continuous.
- Gradient of f being Lipschitz continuous implies

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$

Smooth Functions

- A function f is called smooth if its gradient is Lipschitz continuous.
- Gradient of f being Lipschitz continuous implies

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$

- This also implies that the Hessian satisfies

$$\nabla^2 f(x) \preceq LI$$

Smooth Functions

- A function f is called smooth if its gradient is Lipschitz continuous.
- Gradient of f being Lipschitz continuous implies

$$||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$$

- This also implies that the Hessian satisfies

$$\nabla^2 f(x) \preceq LI$$

- Lemma: If a convex function f is smooth (i.e. has Lipschitz continuous gradients) then:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

Smooth Functions Examples

- Lets consider $f(x) = x^2$. Note that $\nabla f(x) = 2x$.

Smooth Functions Examples

- Lets consider $f(x) = x^2$. Note that $\nabla f(x) = 2x$.
- The condition of Lipschitz smoothness asks for the existence of a L such that

$$||\nabla f(x) - \nabla f(y) \leq L||x - y||$$

Smooth Functions Examples

- Lets consider $f(x) = x^2$. Note that $\nabla f(x) = 2x$.
- The condition of Lipschitz smoothness asks for the existence of a L such that
$$||\nabla f(x) - \nabla f(y) \leq L||x - y||$$
- In this case, this implies $2|y - x| \leq L|y - x|$ which means that $f(x) = x^2$ is Lipschitz continuous for any $L \geq 2$.

Smooth Functions Examples

- Lets consider $f(x) = x^2$. Note that $\nabla f(x) = 2x$.
- The condition of Lipschitz smoothness asks for the existence of a L such that
$$||\nabla f(x) - \nabla f(y) \leq L||x - y||$$
- In this case, this implies $2|y - x| \leq L|y - x|$ which means that $f(x) = x^2$ is Lipschitz continuous for any $L \geq 2$.
- Is $f(x) = x^4$ Lipschitz smooth? What about $f(x) = x^3/3$? Lets study the latter. $\nabla f(x) = x^2$. Lipschitz continuity implies does there exists a L such that $|x^2 - y^2| \leq L|x - y|$ which implies $|x + y| \leq L$. This means that globally, this is not Lipschitz continuous though it can be locally!

Smooth Functions Examples

- Lets consider $f(x) = x^2$. Note that $\nabla f(x) = 2x$.
- The condition of Lipschitz smoothness asks for the existence of a L such that
$$||\nabla f(x) - \nabla f(y) \leq L||x - y||$$
- In this case, this implies $2|y - x| \leq L|y - x|$ which means that $f(x) = x^2$ is Lipschitz continuous for any $L \geq 2$.
- Is $f(x) = x^4$ Lipschitz smooth? What about $f(x) = x^3/3$? Lets study the latter. $\nabla f(x) = x^2$. Lipschitz continuity implies does there exists a L such that $|x^2 - y^2| \leq L|x - y|$ which implies $|x + y| \leq L$. This means that globally, this is not Lipschitz continuous though it can be locally!
- Message: Only functions of asymptotically at most quadratic growth can be smooth globally.

Better understanding Lipschitz continuity and smoothness

- **Consider:** $f'(x) = |x|$

Better understanding Lipschitz continuity and smoothness

- **Consider:** $f'(x) = |x|$
- Since $|f'(x) - f'(y)| = ||x| - |y|| \leq |x - y|$,
  f is Lipschitz continuous with $L = 1$

Better understanding Lipschitz continuity and smoothness

- **Consider:** $f'(x) = |x|$
- Since $|f'(x) - f'(y)| = ||x| - |y|| \leq |x - y|$,
  f is Lipschitz continuous with $L = 1$
- However, it is not differentiable everywhere (not at 0). Such functions are called differentiable almost everywhere.

Better understanding Lipschitz continuity and smoothness

- **Consider:** $f'(x) = |x|$
- Since $|f'(x) - f'(y)| = ||x| - |y|| \leq |x - y|$, f is Lipschitz continuous with $L = 1$
- However, it is not differentiable everywhere (not at 0). Such functions are called differentiable almost everywhere.
- Is every sub-quadratic function Lipschitz smooth? Consider $f(x) = |x|^{3/2}$ on a closed set X s.t. $0 \in X$. Is this function smooth? Note that the gradient of $f'(x)$ is ∞ at $x = 0$.

Lipschitz Continuity on closed and bounded functions

- If f is continuously differentiable everywhere, it is also Lipschitz continuous

Lipschitz Continuity on closed and bounded functions

- If f is continuously differentiable everywhere, it is also Lipschitz continuous
- For functions over a closed and bounded subset of $\Re^n$: f is continuous $\supseteq$ f is differentiable (almost everywhere) $=$ f is Lipschitz continuous $\supseteq$ f is continuously differentiable $=$ ∇f is continuous $\supseteq$ ∇f is differentiable (almost everywhere) $=$ f is smooth

Lipschitz Continuity on closed and bounded functions

- If f is continuously differentiable everywhere, it is also Lipschitz continuous
- For functions over a closed and bounded subset of $\Re^n$: f is continuous $\supseteq$ f is differentiable (almost everywhere) $=$ f is Lipschitz continuous $\supseteq$ f is continuously differentiable $=$ ∇f is continuous $\supseteq$ ∇f is differentiable (almost everywhere) $=$ f is smooth
- Recall that a function is Lipschitz continuous if the norm of the (sub)gradient is bounded!

Lipschitz Continuity on closed and bounded functions

- If f is continuously differentiable everywhere, it is also Lipschitz continuous
- For functions over a closed and bounded subset of $\Re^n$: f is continuous $\supseteq$ f is differentiable (almost everywhere) $=$ f is Lipschitz continuous $\supseteq$ f is continuously differentiable $= \nabla f$ is continuous $\supseteq$ ∇f is differentiable (almost everywhere) $=$ f is smooth
- Recall that a function is Lipschitz continuous if the norm of the (sub)gradient is bounded!
- Also it holds that over a closed and bounded subset of $\Re^n$ that f is Lipschitz continuous $\supseteq$ f is convex

More reading on Lipschitz Continuity

- Juha Heinonen, *Lectures on Lipschitz Analysis*, `http://www.math.jyu.fi/research/reports/rep100.pdf`
- `https://ljk.imag.fr/membres/Anatoli.Iouditski/cours/convex/chapitre_3.pdf`
- Wikipedia: `https://en.wikipedia.org/wiki/Lipschitz_continuity`
- Nice Blog on Lipschitz Continuity: `https://xingyuzhou.org/blog/notes/Lipschitz-gradient`. The author has a similar blog on Strong Convexity: `http://xingyuzhou.org/blog/notes/strong-convexity`

More on Strong Convexity, Lipschitz Continuity and Smoothness

- Recall that a function f is strongly convex if there exists a $\mu > 0$ such that $f(x) - \mu/2||x||^2$ is convex.

More on Strong Convexity, Lipschitz Continuity and Smoothness

- Recall that a function f is strongly convex if there exists a $\mu > 0$ such that $f(x) - \mu/2||x||^2$ is convex.
- Is there is a similar result for Lipschitz smooth functions?

More on Strong Convexity, Lipschitz Continuity and Smoothness

- Recall that a function f is strongly convex if there exists a $\mu > 0$ such that $f(x) - \mu/2||x||^2$ is convex.
- Is there is a similar result for Lipschitz smooth functions?
- A function f is Lipschitz smooth if there exists a $L > 0$ such that $L/2||x||^2 - f(x)$ is convex!

More on Strong Convexity, Lipschitz Continuity and Smoothness

- Recall that a function f is strongly convex if there exists a $\mu > 0$ such that $f(x) - \mu/2||x||^2$ is convex.
- Is there is a similar result for Lipschitz smooth functions?
- A function f is Lipschitz smooth if there exists a $L > 0$ such that $L/2||x||^2 - f(x)$ is convex!
- In fact there is an interesting duality between the two (more on that later).

More on Strong Convexity, Lipschitz Continuity and Smoothness

- Recall that a function f is strongly convex if there exists a $\mu > 0$ such that $f(x) - \mu/2||x||^2$ is convex.
- Is there is a similar result for Lipschitz smooth functions?
- A function f is Lipschitz smooth if there exists a $L > 0$ such that $L/2||x||^2 - f(x)$ is convex!
- In fact there is an interesting duality between the two (more on that later).
- If a Function f is strongly convex and g is convex, the function $f + g$ is strongly convex.

More on Strong Convexity, Lipschitz Continuity and Smoothness

- Recall that a function f is strongly convex if there exists a $\mu > 0$ such that $f(x) - \mu/2||x||^2$ is convex.
- Is there is a similar result for Lipschitz smooth functions?
- A function f is Lipschitz smooth if there exists a $L > 0$ such that $L/2||x||^2 - f(x)$ is convex!
- In fact there is an interesting duality between the two (more on that later).
- If a Function f is strongly convex and g is convex, the function $f + g$ is strongly convex.
- If a function f_1 is Lipschitz smooth and f_2 is Lipschitz smooth, then $f_1 + f_2$ is also Lipschitz smooth

Properties of ML Loss Functions

- Logistic Loss: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i))$: Lipschitz Smooth (Why?)

Properties of ML Loss Functions

- Logistic Loss: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i))$: Lipschitz Smooth (Why?)
- Hinge Loss: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i \theta^T x_i\}$: Lipschitz Continuous

Properties of ML Loss Functions

- Logistic Loss: $L(\theta) = \sum_{i=1}^n \log(1 + \exp(-y_i\theta^T x_i))$: Lipschitz Smooth (Why?)
- Hinge Loss: $L(\theta) = \sum_{i=1}^n \max\{0, 1 - y_i\theta^T x_i\}$: Lipschitz Continuous
- Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^n -\theta_{y_i}^T x_i + \log(\sum_{c=1}^k \exp(\theta_c^T x_i))\}$: Lipschitz Smooth

Properties of ML Loss Functions

- Logistic Loss: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i\theta^T x_i))$: Lipschitz Smooth (Why?)
- Hinge Loss: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i\theta^T x_i\}$: Lipschitz Continuous
- Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\}$: Lipschitz Smooth
- Least Squares: $L(\theta) = \sum_{i=1}^{n}(\theta^T x_i - y_i)^2$: Lipschitz Smooth and Strongly Convex!

Properties of ML Loss Functions

- Logistic Loss: $L(\theta) = \sum_{i=1}^n \log(1 + \exp(-y_i \theta^T x_i))$: Lipschitz Smooth (Why?)
- Hinge Loss: $L(\theta) = \sum_{i=1}^n \max\{0, 1 - y_i \theta^T x_i\}$: Lipschitz Continuous
- Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^n -\theta_{y_i}^T x_i + \log(\sum_{c=1}^k \exp(\theta_c^T x_i))\}$: Lipschitz Smooth
- Least Squares: $L(\theta) = \sum_{i=1}^n (\theta^T x_i - y_i)^2$: Lipschitz Smooth and Strongly Convex!
- L2 Regularization: Lipschitz Smooth and Strongly Convex!

Properties of ML Loss Functions

- Logistic Loss: $L(\theta) = \sum_{i=1}^{n} \log(1 + \exp(-y_i \theta^T x_i))$: Lipschitz Smooth (Why?)
- Hinge Loss: $L(\theta) = \sum_{i=1}^{n} \max\{0, 1 - y_i \theta^T x_i\}$: Lipschitz Continuous
- Multi-class Logistic Regression: $L(\theta_1, \cdots, \theta_k) = \sum_{i=1}^{n} -\theta_{y_i}^T x_i + \log(\sum_{c=1}^{k} \exp(\theta_c^T x_i))\}$: Lipschitz Smooth
- Least Squares: $L(\theta) = \sum_{i=1}^{n} (\theta^T x_i - y_i)^2$: Lipschitz Smooth and Strongly Convex!
- L2 Regularization: Lipschitz Smooth and Strongly Convex!
- L1 Regularization: Lipschitz Continuous

Properties of ML Loss Functions

- L1 Regularized Logistic Loss: Lipschitz Continuous

Properties of ML Loss Functions

- L1 Regularized Logistic Loss: Lipschitz Continuous
- L2 Regularized Logistic Loss: Strongly Convex and Lipschitz Smooth!

Properties of ML Loss Functions

- L1 Regularized Logistic Loss: Lipschitz Continuous
- L2 Regularized Logistic Loss: Strongly Convex and Lipschitz Smooth!
- L1 Regularized Hinge Loss: Lipschitz Continuous

Properties of ML Loss Functions

- L1 Regularized Logistic Loss: Lipschitz Continuous
- L2 Regularized Logistic Loss: Strongly Convex and Lipschitz Smooth!
- L1 Regularized Hinge Loss: Lipschitz Continuous
- L2 Regularized Hinge Loss: Lipschitz Continuous and Strongly Convex (On a Bounded set)

Properties of ML Loss Functions

- L1 Regularized Logistic Loss: Lipschitz Continuous
- L2 Regularized Logistic Loss: Strongly Convex and Lipschitz Smooth!
- L1 Regularized Hinge Loss: Lipschitz Continuous
- L2 Regularized Hinge Loss: Lipschitz Continuous and Strongly Convex (On a Bounded set)
- L2 Regularized Least Squares (Lasso): Lipschitz Smooth and Strongly Convex!

Properties of ML Loss Functions

- L1 Regularized Logistic Loss: Lipschitz Continuous
- L2 Regularized Logistic Loss: Strongly Convex and Lipschitz Smooth!
- L1 Regularized Hinge Loss: Lipschitz Continuous
- L2 Regularized Hinge Loss: Lipschitz Continuous and Strongly Convex (On a Bounded set)
- L2 Regularized Least Squares (Lasso): Lipschitz Smooth and Strongly Convex!
- L1 Regularized Least Squares: Lipschitz Continuous and Strongly Convex (bounded set)

Gradient Descent

- Goal: Given a convex function f, find a $x \in \mathbb{R}^n$ such that $|f(x) - f(x^*)| \leq \epsilon$

Gradient Descent

- Goal: Given a convex function f, find a $x \in \mathbb{R}^n$ such that $|f(x) - f(x^*)| \leq \epsilon$
- Iterative algorithm: Initialize x_0 either randomly or say, 0. Then set

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

γ is the step size parameter which needs to be set.

Gradient Descent

- Goal: Given a convex function f, find a $x \in \mathbb{R}^n$ such that $|f(x) - f(x^*)| \leq \epsilon$
- Iterative algorithm: Initialize x_0 either randomly or say, 0. Then set

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

  γ is the step size parameter which needs to be set.
- Critical Question: How much time does it take to reach an ϵ-approximate solution?

Gradient Descent

- Goal: Given a convex function f, find a $x \in \mathbb{R}^n$ such that $|f(x) - f(x^*)| \leq \epsilon$
- Iterative algorithm: Initialize x_0 either randomly or say, 0. Then set

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

  γ is the step size parameter which needs to be set.
- Critical Question: How much time does it take to reach an ϵ-approximate solution?
- Define x^* as the Global minimizer of f

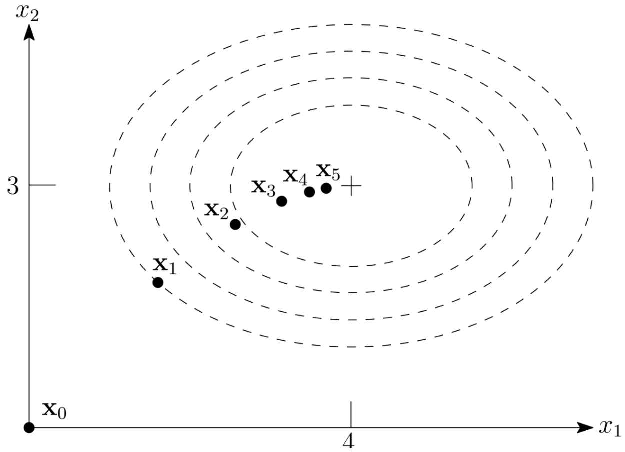
Gradient Descent

- Goal: Given a convex function f, find a $x \in \mathbb{R}^n$ such that $|f(x) - f(x^*)| \leq \epsilon$
- Iterative algorithm: Initialize x_0 either randomly or say, 0. Then set

$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

  γ is the step size parameter which needs to be set.
- Critical Question: How much time does it take to reach an ϵ-approximate solution?
- Define x^* as the Global minimizer of f
- Let f be Lipschitz continuous with parameter B. If f is smooth, let ∇f be Lipschitz continuous with parameter L.

Gradient Descent Illustration



Source: Martin Jaggi (CS 439)

Acknowledgements

In the following slides, I heavily borrow from the notes of Sebastian Bubeck and the slides of Martin Jaggi (EPFL).

Analysis I

- Define $g_t = \nabla f(x_t)$. From the definition of GD:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

Analysis I

- Define $g_t = \nabla f(x_t)$. From the definition of GD:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

- Note that $2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2$

Analysis I

- Define $g_t = \nabla f(x_t)$. From the definition of GD:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

- Note that $2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2$
- We can then rewrite the RHS as:

$$\begin{aligned} g_t^T(x_t - x^*) &= 1/2\gamma(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \\ &= \gamma/2||g_t||^2 + 1/2\gamma(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \end{aligned}$$

Analysis I

- Define $g_t = \nabla f(x_t)$. From the definition of GD:

$$g_t^T(x_t - x^*) = \frac{1}{\gamma}(x_t - x_{t+1})^T(x_t - x^*)$$

- Note that $2v^T w = ||v||^2 + ||w||^2 - ||v - w||^2$
- We can then rewrite the RHS as:

$$\begin{aligned} g_t^T(x_t - x^*) &= 1/2\gamma(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \\ &= \gamma/2||g_t||^2 + 1/2\gamma(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \end{aligned}$$

- Summing this up over

$$\begin{aligned} g_t^T(x_t - x^*) &= 1/2\gamma(||x_t - x_{t+1}||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \\ &= \gamma/2||g_t||^2 + 1/2\gamma(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) \end{aligned}$$

t iterations:

$$\sum_{t=1}^{T-1} g_t^T(x_t - x^*) = \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2 - ||x_T - x^*||^2)$$

Analysis II

- Lets invoke convexity with $x = x_t, y = x^*$.

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*)$$

Analysis II

- Lets invoke convexity with $x = x_t, y = x^*$.

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*)$$

- Recall from the previous slide:

$$\sum_{t=1}^{T-1} g_t^T(x_t - x^*) = \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2 - ||x_T - x^*||^2)$$

which provides:

$$\sum_{t=1}^{T-1} g_t^T(x_t - x^*) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

Analysis II

- Lets invoke convexity with $x = x_t, y = x^*$.

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*)$$

- Recall from the previous slide:

$$\sum_{t=1}^{T-1} g_t^T(x_t - x^*) = \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2 - ||x_T - x^*||^2)$$

which provides:

$$\sum_{t=1}^{T-1} g_t^T(x_t - x^*) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

- Combining both, we have:

$$\sum_{t=0}^{T-1} (f(x_t) - f(x^*)) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

Lipschitz Continuous Functions

- Recall final result:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

Lipschitz Continuous Functions

- Recall final result:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

- Let $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B$ for all x. Set $\gamma = \frac{R}{B\sqrt{T}}$.

Lipschitz Continuous Functions

- Recall final result:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

- Let $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B$ for all x. Set $\gamma = \frac{R}{B\sqrt{T}}$.
- We obtain:

$$\frac{1}{T}\sum_{t=1}^{T-1}[f(x_t) - f(x^*)] \leq \frac{RB}{\sqrt{T}}$$

Lipschitz Continuous Functions

- Recall final result:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

- Let $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B$ for all x. Set $\gamma = \frac{R}{B\sqrt{T}}$.
- We obtain:

$$\frac{1}{T}\sum_{t=1}^{T-1}[f(x_t) - f(x^*)] \leq \frac{RB}{\sqrt{T}}$$

- Last iterate not necessarily the best!

Lipschitz Continuous Functions

- Recall final result:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \gamma/2 \sum_{t=1}^{T-1} ||g_t||^2 + 1/2\gamma(||x_0 - x^*||^2)$$

- Let $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B$ for all x. Set $\gamma = \frac{R}{B\sqrt{T}}$.
- We obtain:

$$\frac{1}{T}\sum_{t=1}^{T-1}[f(x_t) - f(x^*)] \leq \frac{RB}{\sqrt{T}}$$

- Last iterate not necessarily the best!
- Choose $\hat{x} = \text{argmin}_i f(x_i)$ as the final iterate. Show that $|f(\hat{x}) - f(x^*)|$ satisfies the above bound (**exercise**).

Lipschitz Continuous Functions: Final Bound

- Define $\hat{x} = \text{argmin}_i f(x_i)$. Then,

$$|f(\hat{x}) - f(x^*)| \leq \frac{RB}{\sqrt{T}}$$

Lipschitz Continuous Functions: Final Bound

- Define $\hat{x} = \text{argmin}_i f(x_i)$. Then,

$$|f(\hat{x}) - f(x^*)| \leq \frac{RB}{\sqrt{T}}$$

- If we need $|f(\hat{x}) - f(x^*)| \leq \epsilon$, this implies

$$\frac{RB}{\sqrt{T}} \leq \epsilon$$

Lipschitz Continuous Functions: Final Bound

- Define $\hat{x} = \text{argmin}_i f(x_i)$. Then,

$$|f(\hat{x}) - f(x^*)| \leq \frac{RB}{\sqrt{T}}$$

- If we need $|f(\hat{x}) - f(x^*)| \leq \epsilon$, this implies

$$\frac{RB}{\sqrt{T}} \leq \epsilon$$

- Which implies:

$$T \geq \frac{R^2 B^2}{\epsilon^2}$$

Lipschitz Continuous Functions: Final Bound

- Define $\hat{x} = \text{argmin}_i f(x_i)$. Then,

$$|f(\hat{x}) - f(x^*)| \leq \frac{RB}{\sqrt{T}}$$

- If we need $|f(\hat{x}) - f(x^*)| \leq \epsilon$, this implies

$$\frac{RB}{\sqrt{T}} \leq \epsilon$$

- Which implies:

$$T \geq \frac{R^2 B^2}{\epsilon^2}$$

- Final Result: Given a Lipschitz continuous function f, Gradient descent with step size $\gamma = \frac{R}{B\sqrt{T}}$ achieves a solution $\hat{x}$ s.t $|f(\hat{x}) - f(x^*)| \leq \epsilon$ in $\frac{R^2 B^2}{\epsilon^2}$ iterations.

How good or bad is this bound?

- Final Result: Given a B-Lipschitz continuous function convex f, Gradient descent with step size $\gamma = \frac{R}{B\sqrt{T}}$ achieves a solution $\hat{x}$ s.t $|f(\hat{x}) - f(x^*)| \leq \epsilon$ in $\frac{R^2B^2}{\epsilon^2}$ iterations.

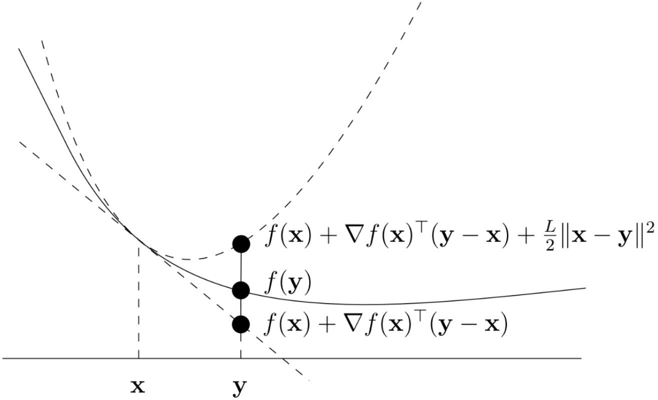
How good or bad is this bound?

- Final Result: Given a B-Lipschitz continuous function convex f, Gradient descent with step size $\gamma = \frac{R}{B\sqrt{T}}$ achieves a solution $\hat{x}$ s.t $|f(\hat{x}) - f(x^*)| \leq \epsilon$ in $\frac{R^2 B^2}{\epsilon^2}$ iterations.
- Advantages of this bound: a) Goes to zero as T gets large, and b) Dimension Independent!

How good or bad is this bound?

- Final Result: Given a B-Lipschitz continuous function convex f, Gradient descent with step size $\gamma = \frac{R}{B\sqrt{T}}$ achieves a solution $\hat{x}$ s.t $|f(\hat{x}) - f(x^*)| \leq \epsilon$ in $\frac{R^2B^2}{\epsilon^2}$ iterations.
- Advantages of this bound: a) Goes to zero as T gets large, and b) Dimension Independent!
- Disadvantages: Slow convergence. To achieve a an error of 0.01, we require $10^4 R^2 B^2$ iterations. To achieve an error of 0.0001, the number of iterations is $10^8 R^2 B^2$!

Smooth Functions



Source: Martin Jaggi (CS 439)

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f
- Properties of L

Recap: Smoothness vs Continuity

- Bounded gradients $\Longleftrightarrow$ Lipschitz continuous f
- Smoothness $\Longleftrightarrow$ Lipschitz continuity of ∇f
- Properties of L
  - Let $f_1, \cdots, f_m$ be smooth convex functions with parameters $L_1, \cdots, L_m$ and let $\lambda_1, \cdots, \lambda_m \geq 0$ be scalars. Then the convex function $f = \sum_{i=1}^m \lambda_i f_i$ is smooth with parameters $\sum +i\lambda_i L_i$

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f
- Properties of L
  - Let $f_1, \cdots, f_m$ be smooth convex functions with parameters $L_1, \cdots, L_m$ and let $\lambda_1, \cdots, \lambda_m \geq 0$ be scalars. Then the convex function $f = \sum_{i=1}^m \lambda_i f_i$ is smooth with parameters $\sum + i\lambda_i L_i$
  - Let f be convex and smooth with parameter L and let $g(x) = Ax + b$ be a vector valued function. Then the convex function $f(g(x))$ is smooth with parameter $L||A||^2 = L\lambda_{\max}(A^T A)$. Here $||A||$ is the spectral norm of A.

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f
- Properties of L
  - Let $f_1, \cdots, f_m$ be smooth convex functions with parameters $L_1, \cdots, L_m$ and let $\lambda_1, \cdots, \lambda_m \geq 0$ be scalars. Then the convex function $f = \sum_{i=1}^{m} \lambda_i f_i$ is smooth with parameters $\sum + i\lambda_i L_i$
  - Let f be convex and smooth with parameter L and let $g(x) = Ax + b$ be a vector valued function. Then the convex function $f(g(x))$ is smooth with parameter $L||A||^2 = L\lambda_{\max}(A^T A)$. Here $||A||$ is the spectral norm of A.
  - Can you use this to derive a bound on the value of L for ∇f where f is the Logistic Loss?

Recap: Smoothness vs Continuity

- Bounded gradients $\iff$ Lipschitz continuous f
- Smoothness $\iff$ Lipschitz continuity of ∇f
- Properties of L
  - Let $f_1, \cdots, f_m$ be smooth convex functions with parameters $L_1, \cdots, L_m$ and let $\lambda_1, \cdots, \lambda_m \geq 0$ be scalars. Then the convex function $f = \sum_{i=1}^m \lambda_i f_i$ is smooth with parameters $\sum + i\lambda_i L_i$
  - Let f be convex and smooth with parameter L and let $g(x) = Ax + b$ be a vector valued function. Then the convex function $f(g(x))$ is smooth with parameter $L||A||^2 = L\lambda_{\max}(A^T A)$. Here $||A||$ is the spectral norm of A.
  - Can you use this to derive a bound on the value of L for ∇f where f is the Logistic Loss?
- Recall:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

Gradient Descent for Smooth Functions: I

- Recall:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

Gradient Descent for Smooth Functions: I

- Recall:
$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$
- Note $x_{t+1} - x_t = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

$$\begin{aligned} f(x_{t+1} &\leq f(x_t) + g_t^T(x_{t+1} - x_t) + L/2||x_{t+1} - x_t||^2 && (1) \\ &\leq f(x_t) - \gamma||g_t||^2 + L/2\gamma^2||g_t||^2 && (2) \end{aligned}$$

Gradient Descent for Smooth Functions: I

- Recall:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

- Note $x_{t+1} - x_t = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

$$f(x_{t+1} \leq f(x_t) + g_t^T(x_{t+1} - x_t) + L/2||x_{t+1} - x_t||^2 \quad (1)$$
$$\leq f(x_t) - \gamma||g_t||^2 + L/2\gamma^2||g_t||^2 \quad (2)$$

- Set the step size as $\gamma = 1/L$. The above result becomes:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2$$

Gradient Descent for Smooth Functions: I

- Recall:

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}||\mathbf{y} - \mathbf{x}||^2$$

- Note $x_{t+1} - x_t = -\gamma g_t$. Also substituting $y = x_{t+1}$ and $x = x_t$ above and doing some math, we obtain

$$f(x_{t+1} \leq f(x_t) + g_t^T(x_{t+1} - x_t) + L/2||x_{t+1} - x_t||^2 \quad (1)$$
$$\leq f(x_t) - \gamma||g_t||^2 + L/2\gamma^2||g_t||^2 \quad (2)$$

- Set the step size as $\gamma = 1/L$. The above result becomes:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2$$

- This means GD is guaranteed to decrease the function value at every iteration!

Gradient Descent for Smooth Functions: II

- Lets start with:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2 \Rightarrow \frac{1}{2L}||g_t||^2 \leq f(x_t) - f(x_{t+1})$$

Gradient Descent for Smooth Functions: II

- Lets start with:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2 \Rightarrow \frac{1}{2L}||g_t||^2 \leq f(x_t) - f(x_{t+1})$$

- Summing this from $t = 1$ to $T - 1$:

$$\frac{1}{2L}\sum_{t=1}^{T-1}||g_t||^2 \leq \sum_{t=1}^{T-1}(f(x_t) - f(x_{t+1}) = f(x_0) - f(x_T)$$

Gradient Descent for Smooth Functions: II

- Lets start with:

$$f(x_{t+1}) \leq f(x_t) - \frac{1}{2L}||g_t||^2 \Rightarrow \frac{1}{2L}||g_t||^2 \leq f(x_t) - f(x_{t+1})$$

- Summing this from $t = 1$ to $T - 1$:

$$\frac{1}{2L}\sum_{t=1}^{T-1}||g_t||^2 \leq \sum_{t=1}^{T-1}(f(x_t) - f(x_{t+1}) = f(x_0) - f(x_T)$$

- Next, recall from Analysis II (and after setting $\gamma = 1/L$

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{1}{2L}\sum_{t=1}^{T-1}||g_t||^2 + \frac{L}{2}(||x_0 - x^*||^2)$$
$$\leq f(x_0) - f(x_T) + \frac{L}{2}||x_0 - x^*||^2$$

Gradient Descent for Smooth Functions: III

- We had:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq f(x_0) - f(x_T) + \frac{L}{2}||x_0 - x^*||^2$$

Gradient Descent for Smooth Functions: III

- We had:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq f(x_0) - f(x_T) + \frac{L}{2}||x_0 - x^*||^2$$

- Re-writing the math:

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2$$

Gradient Descent for Smooth Functions: III

- We had:

$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq f(x_0) - f(x_T) + \frac{L}{2}||x_0 - x^*||^2$$

- Re-writing the math:

$$\sum_{t=1}^{T}(f(x_t) - f(x^*)) \leq \frac{L}{2}||x_0 - x^*||^2$$

- This implies that (**why?**):

$$f(x_T) - f(x^*) \leq \sum_{t=1}^{T}\frac{(f(x_t) - f(x^*))}{T} \leq \frac{L}{2T}||x_0 - x^*||^2$$

Convergence rate for Smooth Functions

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$

Convergence rate for Smooth Functions

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$
- To ensure that $f(x_T) - f(x^*) \leq \epsilon$, we require $\frac{LR^2}{2T} \leq \epsilon$.

Convergence rate for Smooth Functions

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$
- To ensure that $f(x_T) - f(x^*) \leq \epsilon$, we require $\frac{LR^2}{2T} \leq \epsilon$.
- This implies that $T \geq \frac{R^2 L}{2\epsilon}$

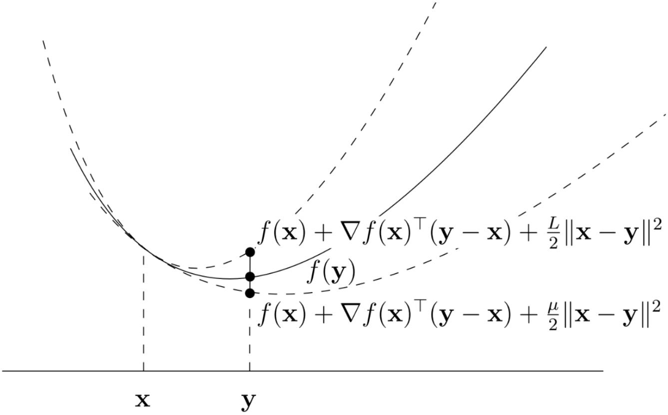
Convergence rate for Smooth Functions

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$
- To ensure that $f(x_T) - f(x^*) \leq \epsilon$, we require $\frac{LR^2}{2T} \leq \epsilon$.
- This implies that $T \geq \frac{R^2 L}{2\epsilon}$
- To achieve an error of 0.01, we require $50R^2L$ iterations instead of $10^4 R^2 B^2$ in the Lipschitz case!

Convergence rate for Smooth Functions

- Putting everything together: $f(x_T) - f(x^*) \leq \frac{L}{2T}||x_0 - x^*||^2 = \frac{LR^2}{2T}$
- To ensure that $f(x_T) - f(x^*) \leq \epsilon$, we require $\frac{LR^2}{2T} \leq \epsilon$.
- This implies that $T \geq \frac{R^2L}{2\epsilon}$
- To achieve an error of 0.01, we require $50R^2L$ iterations instead of $10^4R^2B^2$ in the Lipschitz case!
- Final Result: Given a L smooth convex function f, Gradient descent with step size $\gamma = \frac{1}{L}$ achieves a solution x_T s.t $|f(x_T) - f(x^*)| \leq \epsilon$ in $\frac{R^2L}{\epsilon}$ iterations.

Smooth + Strongly Convex Functions



Source: Martin Jaggi (CS 439)

Fastest Convergence with Smooth + Strongly Convex I

- Recall from Analysis I:

$$g_t^T(x_t - x^*) = \gamma_t/2||g_t||^2 + 1/2\gamma_t(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)$$

Fastest Convergence with Smooth + Strongly Convex I

- Recall from Analysis I:

$$g_t^T(x_t - x^*) = \gamma_t/2||g_t||^2 + 1/2\gamma_t(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)$$

- We can use a stronger lower bound on the LHS via strong convexity: $g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2$

Fastest Convergence with Smooth + Strongly Convex I

- Recall from Analysis I:

$$g_t^T(x_t - x^*) = \gamma_t/2||g_t||^2 + 1/2\gamma_t(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)$$

- We can use a stronger lower bound on the LHS via strong convexity: $g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2$
- Putting both together and next rearranging terms:

$$f(x_t) - f(x^*) \leq \frac{1}{2\gamma}(\gamma^2||g_t||^2 + ||x_t - x^*||^2 - ||x_{t+1} - x^*||^2) - \frac{\mu}{2}||x_t - x^*||^2$$

$$\Rightarrow ||x_{t+1} - x^*||^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex II

- From previous slide:

$$||x_{t+1} - x^*||^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex II

- From previous slide:

$$||x_{t+1} - x^*||^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

- Now lets show that $2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 \leq 0$. Lets set the step size $\gamma = 1/L$.

$$\begin{aligned} 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 &\leq \frac{2}{L}(f(x_{t+1}) - f(x_t)) + \frac{1}{L^2}||g_t||^2 \\ &\leq -\frac{1}{L^2}||g_t||^2 + \frac{1}{L^2}||g_t||^2 = 0 \end{aligned}$$

Fastest Convergence with Smooth + Strongly Convex II

- From previous slide:

$$||x_{t+1} - x^*||^2 \leq 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

- Now lets show that $2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 \leq 0$. Lets set the step size $\gamma = 1/L$.

$$\begin{aligned} 2\gamma(f(x^*) - f(x_t)) + \gamma^2||g_t||^2 &\leq \frac{2}{L}(f(x_{t+1}) - f(x_t)) + \frac{1}{L^2}||g_t||^2 \\ &\leq -\frac{1}{L^2}||g_t||^2 + \frac{1}{L^2}||g_t||^2 = 0 \end{aligned}$$

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq (1 - \frac{\mu}{L})||x_t - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq (1 - \frac{\mu}{L})||x_t - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq (1 - \frac{\mu}{L})||x_t - x^*||^2$$

- Multiplying all terms from $t = 0, \cdots, T-1$:

$$x_T - x^*||^2 \leq (1 - \frac{\mu}{L})^T ||x_0 - x^*||^2$$

Fastest Convergence with Smooth + Strongly Convex III

- Packing everything together:

$$||x_{t+1} - x^*||^2 \leq (1 - \frac{\mu}{L})||x_t - x^*||^2$$

- Multiplying all terms from $t = 0, \cdots, T-1$:

$$x_T - x^*||^2 \leq (1 - \frac{\mu}{L})^T ||x_0 - x^*||^2$$

- Final step: Lets combine smoothness and the fact that $\nabla f(x^*) = 0$:

$$f(x_T) - f(x^*) \leq \nabla f(x^*)^T (x_T - x^*) + \frac{L}{2}||x_T - x^*||^2 = \frac{L}{2}||x_T - x^*||^2$$

$$\Rightarrow f(x_T) - f(x^*) \leq \frac{L}{2}||x_T - x^*||^2 \leq \frac{L}{2}(1 - \frac{\mu}{L})^T ||x_0 - x^*||^2$$

Convergence Rate For Smooth + Strongly Convex

- Set $R^2 = ||x_0 - x^*||^2$. We get:

$$f(x_T) - f(x^*) \leq \frac{L}{2}(1 - \frac{\mu}{L})^T R^2$$

Convergence Rate For Smooth + Strongly Convex

- Set $R^2 = ||x_0 - x^*||^2$. We get:

$$f(x_T) - f(x^*) \leq \frac{L}{2}(1 - \frac{\mu}{L})^T R^2$$

- To get an error of ϵ, we require $\frac{L}{2}(1 - \frac{\mu}{L})^T R^2 \leq \epsilon$ which implies $T \geq \frac{L}{\mu} \log(\frac{R^2 L}{2\epsilon})$.

Convergence Rate For Smooth + Strongly Convex

- Set $R^2 = ||x_0 - x^*||^2$. We get:

$$f(x_T) - f(x^*) \leq \frac{L}{2}(1 - \frac{\mu}{L})^T R^2$$

- To get an error of ϵ, we require $\frac{L}{2}(1 - \frac{\mu}{L})^T R^2 \leq \epsilon$ which implies $T \geq \frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$.
- To get an error of $\epsilon = 0.01$, we now need only $L/\mu \log(50R^2 L)$ iterations as opposed to $50R^2 L$ iterations in the smooth case!

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in $R^2 B^2/\epsilon^2$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in $R^2 B^2/\epsilon^2$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2 L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, S = 50, SS = 8.49 iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, S = 50, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, S = 500, SS = 13.49 iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in $R^2 B^2/\epsilon^2$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2 L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, S $= 50$, SS $= 8.49$ iterations
  - $\epsilon = 0.01$, C: 1000000, S $= 500$, SS $= 13.49$ iterations
  - $\epsilon = 0.001$, C: 100000000, S $= 5000$, SS $= 18.49$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, S $=$ 50, SS $=$ 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, S $=$ 500, SS $=$ 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000, S $=$ 5000, SS $=$ 18.49 iterations
- As ϵ reduces by 10, the number of iterations of strongly + smooth case increases only by a additive constant! This is linear convergence!

Can we better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?

Can we better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!

Can we better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)

Can we better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)
- Can we get an improved convergence rate in such a case?

Can we better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)
- Can we get an improved convergence rate in such a case?
- We can obtain an improved $O(1/\epsilon)$ bound!!

Lipschitz + Strongly Convex I

- Assume the gradients $||g_t|| \leq B$.

Lipschitz + Strongly Convex I

- Assume the gradients $||g_t|| \leq B$.
- Recall Analysis I:

$$g_t^T(x_t - x^*) = \gamma_t/2||g_t||^2 + 1/2\gamma_t(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)$$

Lipschitz + Strongly Convex I

- Assume the gradients $||g_t|| \leq B$.
- Recall Analysis I:

$$g_t^T(x_t - x^*) = \gamma_t/2||g_t||^2 + 1/2\gamma_t(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2)$$

- Combining with strong convexity we get (after $||g_t|| \leq B^2$)

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + (\gamma_t^{-1} - \mu)/2||x_t - x^*|| - \gamma_t^{-1}/2||x_{t+1} - x^*||^2$$

Lipschitz + Strongly Convex II

- So Far:

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + (\gamma_t^{-1} - \mu)/2||x_t - x^*|| - \gamma_t^{-1}/2||x_{t+1} - x^*||^2$$

Lipschitz + Strongly Convex II

- So Far:

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + (\gamma_t^{-1} - \mu)/2||x_t - x^*|| - \gamma_t^{-1}/2||x_{t+1} - x^*||^2$$

- Plug in $\gamma_t^{-1} = \mu(1+t)/2$ and multily by t on both sides:

$$\begin{aligned} t[f(x_t) - f(x^*)] &\leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\{t(t-1)||x_t - x^*||^2 - (t+1)t||x_{t+1} - x^*||^2\} \\ &\leq \frac{B^2}{\mu} + \frac{\mu}{4}\{t(t-1)||x_t - x^*||^2 - (t+1)t||x_{t+1} - x^*||^2\} \end{aligned}$$

Lipschitz + Strongly Convex II

- So Far:

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + (\gamma_t^{-1} - \mu)/2||x_t - x^*|| - \gamma_t^{-1}/2||x_{t+1} - x^*||^2$$

- Plug in $\gamma_t^{-1} = \mu(1+t)/2$ and multily by t on both sides:

$$\begin{aligned} t[f(x_t) - f(x^*)] &\leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\{t(t-1)||x_t - x^*||^2 - (t+1)t||x_{t+1} - x^*||^2\} \\ &\leq \frac{B^2}{\mu} + \frac{\mu}{4}\{t(t-1)||x_t - x^*||^2 - (t+1)t||x_{t+1} - x^*||^2\} \end{aligned}$$

- Now we can use the telescoping sum and obtain...

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq \frac{TB^2}{\mu} + \mu/4(0 - T(T+1)||x_{T+1} - x^*||^2) \leq TB^2/\mu$$

Lipschitz + Strongly Convex III

- So Far:

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq TB^2/\mu$$

Lipschitz + Strongly Convex III

- So Far:

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq TB^2/\mu$$

- Multiply by $2/(T(T+1))$ on both sides to make it a convex combination:

$$\sum_{t=1}^{T} \frac{2t}{T(T+1)}(f(x_t) - f(x^*)) \leq \frac{2B^2}{\mu(T+1)}$$

Lipschitz + Strongly Convex III

- So Far:

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq TB^2/\mu$$

- Multiply by $2/(T(T+1))$ on both sides to make it a convex combination:

$$\sum_{t=1}^{T} \frac{2t}{T(T+1)}(f(x_t) - f(x^*)) \leq \frac{2B^2}{\mu(T+1)}$$

- This implies if the error $\leq \epsilon$ implies $T \geq \frac{2B^2}{\mu\epsilon} - 1$

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in $R^2 B^2/\epsilon^2$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in $R^2 B^2/\epsilon^2$ iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2 L}{\epsilon}$ iterations.

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, S = 50, SS = 8.49 iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, S $=$ 50, SS $=$ 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, CS: 20000, S $=$ 500, SS $=$ 13.49 iterations

Summary of Results so Far...

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS) with $\gamma_t = \mu(1+t)/2$ achieve an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, S = 50, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, CS: 20000, S = 500, SS = 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000, CS: 200000, S = 5000, SS = 18.49 iterations

Lower Bounds (No Proof)

- Case I: Lipschitz Continuous: Any black-box procedure will have an error of at least $\frac{RB}{2(1+\sqrt{t})}$ (GD: $\frac{RB}{\sqrt{T}}$)

Lower Bounds (No Proof)

- Case I: Lipschitz Continuous: Any black-box procedure will have an error of at least $\frac{RB}{2(1+\sqrt{t})}$ (GD: $\frac{RB}{\sqrt{T}}$)
- Case II: Lipschitz Continuous + Strongly Convex: Any black-box procedure have an error of at least $\frac{B^2}{2\mu T}$, (GD: $\frac{2B^2}{\mu(T+1)}$)

Lower Bounds (No Proof)

- Case I: Lipschitz Continuous: Any black-box procedure will have an error of at least $\frac{RB}{2(1+\sqrt{t})}$ (GD: $\frac{RB}{\sqrt{T}}$)
- Case II: Lipschitz Continuous + Strongly Convex: Any black-box procedure have an error of at least $\frac{B^2}{2\mu T}$, (GD: $\frac{2B^2}{\mu(T+1)}$)
- Case III: Smooth: Any black box procedure have an error of at least $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ (GD: $\frac{LR^2}{2T}$)

Lower Bounds (No Proof)

- Case I: Lipschitz Continuous: Any black-box procedure will have an error of at least $\frac{RB}{2(1+\sqrt{t})}$ (GD: $\frac{RB}{\sqrt{T}}$)
- Case II: Lipschitz Continuous + Strongly Convex: Any black-box procedure have an error of at least $\frac{B^2}{2\mu T}$, (GD: $\frac{2B^2}{\mu(T+1)}$)
- Case III: Smooth: Any black box procedure have an error of at least $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ (GD: $\frac{LR^2}{2T}$)
- Case IV: Smooth + Strongly Convex: Define $\kappa = \frac{L}{\mu}$. Then Any black box procedure will have an error of at least $\frac{\mu}{2}(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1})^{2(T-1)}$ (GD: $\frac{L}{2}(1 - \frac{\mu}{L})^T = \frac{L}{2}(\frac{\kappa-1}{\kappa})^T$)

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.
- Define $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. $\gamma_t \leq 0$

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.
- Define $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. $\gamma_t \leq 0$
- Initialize $x_1 = y_1$ as an arbitrary point

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.
- Define $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. $\gamma_t \leq 0$
- Initialize $x_1 = y_1$ as an arbitrary point
- Step 1: $y_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$ (like normal GD)

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.
- Define $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. $\gamma_t \leq 0$
- Initialize $x_1 = y_1$ as an arbitrary point
- Step 1: $y_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$ (like normal GD)
- Step 2: $x_{t+1} = (1-\gamma_t)y_{t+1} + \gamma_t y_t$ (slide a little bit further than y_{t+1} towards the previous point y_t!)

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.
- Define $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. $\gamma_t \leq 0$
- Initialize $x_1 = y_1$ as an arbitrary point
- Step 1: $y_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$ (like normal GD)
- Step 2: $x_{t+1} = (1-\gamma_t)y_{t+1} + \gamma_t y_t$ (slide a little bit further than y_{t+1} towards the previous point y_t!)
- Theorem (Nesterov 1983): If f is convex and L-smooth, $f(y_T) - f(x^*) \leq \frac{2LR^2}{T^2}$

Nesterov's Accelerated Gradient Descent

- There is a gap of a factor of T for the Smooth case! $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ vs $\frac{LR^2}{2T}$!
- Nesterov's accelerated algorithm fixes this (a very unintuitive algorithm and proof procedure – we will not prove it in the class)
- The algorithm follows the following procedure.
- Define $\lambda_0 = 0, \lambda_t = \frac{1+\sqrt{1+4\lambda_{t-1}^2}}{2}$ and $\gamma_t = \frac{1-\lambda_t}{\lambda_{t+1}}$. $\gamma_t \leq 0$
- Initialize $x_1 = y_1$ as an arbitrary point
- Step 1: $y_{t+1} = x_t - \frac{1}{L}\nabla f(x_t)$ (like normal GD)
- Step 2: $x_{t+1} = (1-\gamma_t)y_{t+1} + \gamma_t y_t$ (slide a little bit further than y_{t+1} towards the previous point y_t!)
- Theorem (Nesterov 1983): If f is convex and L-smooth, $f(y_T) - f(x^*) \leq \frac{2LR^2}{T^2}$
- Matches the lower bound upto contant factors!

Summary of Results so Far...

- Lipschitz continuous functions (C): $R^2 B^2 / \epsilon^2$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions AGD (SAGD): $\sqrt{\frac{2LR^2}{\epsilon}}$ iterations

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions AGD (SAGD): $\sqrt{\frac{2LR^2}{\epsilon}}$ iterations
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions AGD (SAGD): $\sqrt{\frac{2LR^2}{\epsilon}}$ iterations
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions AGD (SAGD): $\sqrt{\frac{2LR^2}{\epsilon}}$ iterations
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, SGD = 50, SAGD = 14.4, SS = 8.49 iterations

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions AGD (SAGD): $\sqrt{\frac{2LR^2}{\epsilon}}$ iterations
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, SGD = 50, SAGD = 14.4, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, CS: 20000, SGD = 500, SAGD: 44.72, SS = 13.49 iterations

Summary of Results so Far...

- Lipschitz continuous functions (C): R^2B^2/ϵ^2 iterations
- Lipschitz continuous functions + Smooth (CS): $2B^2/\epsilon - 1$ iterations
- Smooth Functions GD (SGD): $\frac{R^2L}{\epsilon}$ iterations.
- Smooth Functions AGD (SAGD): $\sqrt{\frac{2LR^2}{\epsilon}}$ iterations
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then we have the:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, SGD = 50, SAGD = 14.4, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, CS: 20000, SGD = 500, SAGD: 44.72, SS = 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000, CS: 200000, SGD = 5000, SAGD = 141.42, SS = 18.49 iterations

Does this Matter in Practice?

- Convergence results and Lower bounds are often worst case!

Does this Matter in Practice?

- Convergence results and Lower bounds are often worst case!
- Though there exists family of functions where the bounds are tight, it is not necessary that the same intuition carries over in practice!

Does this Matter in Practice?

- Convergence results and Lower bounds are often worst case!
- Though there exists family of functions where the bounds are tight, it is not necessary that the same intuition carries over in practice!
- Difference between Theory and Practice and the need to connect the two!

Does this Matter in Practice?

- Convergence results and Lower bounds are often worst case!
- Though there exists family of functions where the bounds are tight, it is not necessary that the same intuition carries over in practice!
- Difference between Theory and Practice and the need to connect the two!
- Next, we will implement some of these algorithms for various ML Loss Functions!

Gradient Descent in Practice: Basic Version

- Credits to Mark Schmidt from UBC for this (I converted his Matlab based tutorial to python)

- Credits to Mark Schmidt from UBC for this (I converted his Matlab based tutorial to python)
- Let us first initialize the Logistic Loss on a dataset

Gradient Descent in Practice: Basic Version

- Credits to Mark Schmidt from UBC for this (I converted his Matlab based tutorial to python)
- Let us first initialize the Logistic Loss on a dataset
- Next, implement a simple gradient descent algorithm.

```
def gd(funObj,w,maxEvals,alpha, ...
X,y,lam,verbosity, freq):
    ...
```

[python]

Gradient Descent in Practice: Basic Version

- Credits to Mark Schmidt from UBC for this (I converted his Matlab based tutorial to python)
- Let us first initialize the Logistic Loss on a dataset
- Next, implement a simple gradient descent algorithm.

```
def gd(funObj,w,maxEvals,alpha, ...
X,y,lam,verbosity, freq):
    ...
```

[python]

  - ‘funObj‘ is the

Gradient Descent in Practice: Basic Version

```
def gd(funObj,w,maxEvals,alpha,X,y,lam, verbosity):
    [f,g] = funObj(w,X,y,lam)
    funEvals = 1
    funVals = []
    while(1):
        [f,g] = funObj(w,X,y,lam)
        optCond = LA.norm(g, np.inf)
        if (verbosity > 0):
            print(funEvals,alpha,f,optCond)
        w = w - alpha*g
        funEvals = funEvals+1
        if ((optCond < 1e-2) and (funEvals > maxEvals)
            break
        funVals.append(f)
    return funVals
```

Gradient Descent in Practice: Basic Version

- Run this by invoking:

```
funV = gd(LogisticLoss,w,200,1e-1,X,y,1,1,10)
```

- Try running this with different values of learning rates: $\alpha = 1e - 1, 1e - 3, 1e - 5, \ldots$
- How do we find the optimal learning rate every time?
- Can there be better strategies to adapt the learning rates?
- Next, we shall see a few line search based strategies.

Armijo Backtracking Line-Search V1

- We don't want to tune α every time
- This is the idea behind line search
- Simple Line search strategy:
  - Start with a large value of α
  - Divide α by 1/2 if it doesn't satisfy Armijo's condition:

$$f(w - \alpha g) \leq f(w) - \gamma\alpha||g||^2$$

  - Basically find α such that there is a reduction in function value by atleast $\gamma\alpha||g||^2$
  - Idea: Choose α and γ such that this happens.

Armijo Backtracking Line-Search V2

- Danger with the simple backtracking is that α may quickly become very small quickly
- Easy fix: Reset α every time!
- Issue with this: Too many function evaluations lost in repeated backtracking!

Armijo Backtracking Line-Search V3

- Just halving the step size ignores the information collected during line search!
- Reduce the number of backtracks using a polynomial interpolation!
- Minimize a quadratic passing through $f(w), f'(w)$ and $f(w - \alpha g)$
- Choose α using a polynomial interpolation as follows:

$$\alpha = \frac{\alpha^2 g^T g}{2(fcurr + \alpha g^T g - f)}$$

- Here *fcurr* is the function evaluation with the current value of α and f is the function value before starting backtracking!

Armijo Backtracking Line-Search V4

- Final Issue to fix is better initialization of α.
- Initializing $\alpha = 1$ is too large in practice
- Wasted backtracks because of this.
- Use a hueristic like $\alpha = 1/||g||$
- On subsequent iterations again use a polynomial interpolation:

$$\alpha = \min(1, 2(f_{old} - f)/g^T g)$$

- A lot of this is tried empirically and based on empirical knowledge..