Optimization in Machine Learning

Lecture 12: Concluding Lipschitz Continuity and Smoothness illustration on ML objectives, Algorithms for Optimization and their analysis

Ganesh Ramakrishnan

Department of Computer Science
Dept of CSE, IIT Bombay
https://www.cse.iitb.ac.in/~ganesh

February, 2025

[Recap] Lipschitz Continuity on closed and bounded functions

- If f is continuously differentiable almost everywhere, it is also Lipschitz continuous
- For functions over a bounded subset of $\Re^n$: f is continuous $\supseteq$ f is differentiable (almost everywhere) $= f$ is Lipschitz continuous[1] $\supseteq$ f is continuously differentiable $= \nabla f$ is continuous $\supseteq \nabla f$ 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

[1] Theorem 1.41 of https://moodle.iitb.ac.in/mod/resource/view.php?id=33825

[Recap] 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

  And so is f1(f2(x)

[Homework] Properties of ML Loss Functions

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

[Homework] Properties of ML Loss Functions

Proof sketch: Using composition

① $f_1(t) = \log(1+e^t)$ $\quad f_1'(t) = e^t/(1+e^t)$ $\quad f_1''(t) = \nabla^2 f(t) = \dfrac{e^t(1+e^t) - e^t \cdot e^t}{(1+e^t)^2}$

$= \dfrac{e^t}{(1+e^t)^2} \overset{?}{\leq} L \quad (L = 1/2 \text{ or } 1/4)$

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

Can be shown to be also Lipschitz Continuous

Using Quotient Rule

$$\frac{d\left(\frac{a(t)}{b(t)}\right)}{dt} = \frac{a'(t)b(t) - b'(t)a(t)}{(b(t))^2}$$

② $f_2(\theta) = -y_i \theta^T x_i$ THe Hessian Is indeed upper bounded by an L = 0

$\sum_i f_1(f_2(\theta))$ would therefore be Lipschitz Smooth

In fact it is also convex

[Homework] Properties of ML Loss Functions

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

[Homework] Properties of ML Loss Functions

Consider simpler function 1) $\max(0, 1-t) = f(t)$

Case 1: $t_1 < 1$ & $t_2 \geq 1$ or $t_1 \geq 1$ & $t_2 < 1$ → Can be done similarly

Case 2: $t_1 < 1$ & $t_2 < 1$

Case 3: $t_1 \geq 1$ & $t_2 \geq 1$

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

Wont expect smoothness (amounts to differentiability of its gradient almost everywhere) Case for Hinge Loss and L1 regularization

Case 1: $|f(t_1) - f(t_2)| = |1 - t_1 - 0| = |1 - t_1| \leq |t_2 - t_1|$ $(L = 1)$

Case 2: $|f(t_1) - f(t_2)| = |(1 - t_1) - (1 - t_2)| = |t_2 - t_1| \leq |t_2 - t_1|$ $(L = 1)$

Case 3: $|f(t_1) - f(t_2)| = 0 \leq |t_2 - t_1|$ $(L \leq 1)$

Subsequently, we consider a composition of the function f(.) above with a linear function exactly as in the previous case

[Homework] Properties of ML Loss Functions

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

[Homework] Properties of ML Loss Functions

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

Proof Methodology: Generalization of the Logistic Loss case

[Homework] Properties of ML Loss Functions

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

[Homework] Properties of ML Loss Functions

The first three are also convex

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

$f(t) = t^2$

Recap this is both Lipschitz Smooth and Strongly Convex

In a bounded domain, we also expect Least Squares to be Lipschitz continuous

[Homework] Properties of ML Loss Functions

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

[Homework] Properties of ML Loss Functions

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

same reasoning

$\|x\|_2^2 = f(\|x\|_2)$

where $f(t) = t^2$

[Homework] Properties of ML Loss Functions

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

[Homework] Properties of ML Loss Functions

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

Recap: We showed that $|x|$ is Lipschitz continuous

[Homework] Properties of ML Loss Functions with Regularizers

- L1 Regularized Logistic Loss: Lipschitz Continuous and Convex

[Homework] Properties of ML Loss Functions with Regularizers

Recap: Sum of Lipschitz continuous functions is Lipschitz continuous
Recap: Sum of convex functions is convex

- L1 Regularized Logistic Loss: Lipschitz Continuous and Convex

[Homework] Properties of ML Loss Functions with Regularizers

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

[Homework] Properties of ML Loss Functions with Regularizers

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

[Homework] Properties of ML Loss Functions with Regularizers

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

[Homework] Properties of ML Loss Functions with Regularizers

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

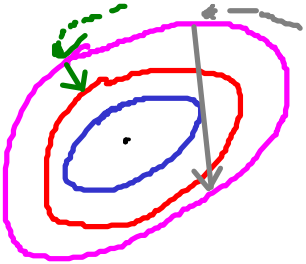
[Homework] Properties of ML Loss Functions with Regularizers

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

Algorithms for Optimization

Algorithms for Optimization

Also has different direction from the gray step



Larger magnitude step (in the Level-Set space)

Two key components of any algorihtm

1) Magnitude
2) Direction

To show/analyze convergence, we would like to assess
1) How fast the function changes (Lipschitz continuity)
2) What is the curvature of the function? Upper bounded? Lower bounded?
3) If the function oscillates (because of drastic changes etc) then updates might also lead to oscilllating solutions
4) Convexity ensures existence global minimum

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

We expect x^* to exist

$x^* = \operatorname{argmin}_x f(x)$

- 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

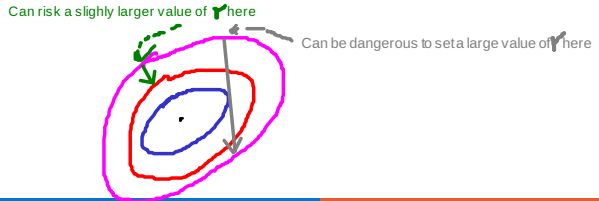
$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

The magnitude part

General Descent Algorithms replace the gradient with some descent step
This component is critical for the Direction component

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

Gradient points in the direction ORTHOGONAL to the level curve at that iterate/point



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?

fn of ϵ

(1) $\frac{1}{\epsilon}$ (2) $\frac{1}{\epsilon^2}$ (3) $\log(\frac{1}{\epsilon})$

Of the three, most preferred will be (3)

(3) > (1) > (2)

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

- 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

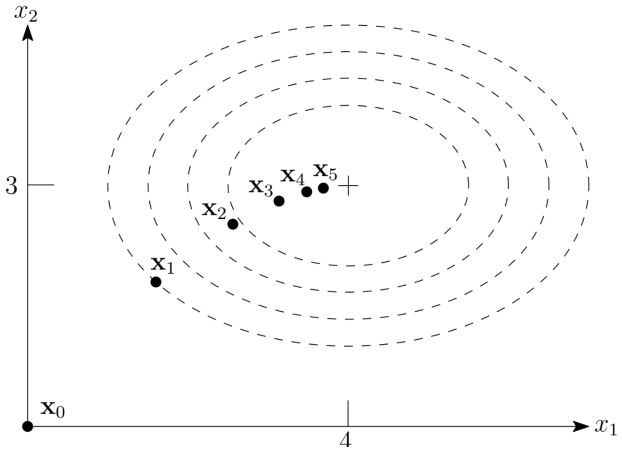
$$x_{t+1} = x_t - \gamma \nabla f(x_t)$$

  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.

We invoke L-continuity since in practice it is more generic (especially on bounded sets)
Interesting: This apparently weaker condition leads to weaker time guarantees!!

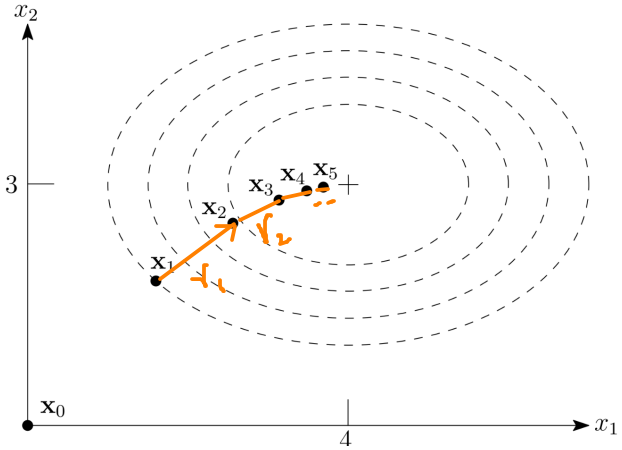
Yields better time guarantees!

Gradient Descent Illustration



Source: Martin Jaggi (CS 439)

Gradient Descent Illustration



Source: Martin Jaggi (CS 439)

Analysis I

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

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

DOT PRODUCT FOR UNSCALED DIRECTIONAL DERIVATIVE ALONG THE DIRECTION FOMR THE CURRENT ITERATE TO THE OPTIMAL SOLUTION

Analysis I

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

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

$x_{t+1} = x_t - \gamma g_t$

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

LHS $2v^T w = \|v\|^2 + \|w\|^2 - \|v-w\|^2$

A trick we will often invoke is transformation of dot products to quadratic expressions

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

$x_{t+1} = x_t - \gamma g_t$

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

LHS $\rightarrow$ $2v^T w = ||v||^2 + ||w||^2 - ||v-w||^2$

A trick we will often invoke is transformation of dot products to quadratic expressions

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:

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

Analysis I

Food for thought: Can we seek such telescopic summations on differences of function values rather than of points?

- 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:

Transform RHS to a summation of differences from x* across iterates

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

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

$$\sum_{t=1}^{T-1} g_t^T(x_t - x^*) = \sum_{t=1}^{T-1} \frac{\gamma}{2}||g_t||^2 + \frac{1}{2\gamma}\left(||x_1 - x^*||^2 - ||x_T - x^*||^2\right)$$

$$||x_1 - x^*||^2 - ||x_2 - x^*||^2 + ||x_2 - x^*||^2 - ||x_3 - x^*||^2 + ||x_3 - \cdots - ||x_T - x^*||^2$$

ILLUSTRATION OF TELESCOPIC SUM

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:

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

- Summing (1) over t iterations:

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

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:

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

Can this be upperbounded based on L-continuity?

- Summing (1) over t iterations:

Can this term be upperbounded based on convexity?

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

Part II: Invoking Convexity

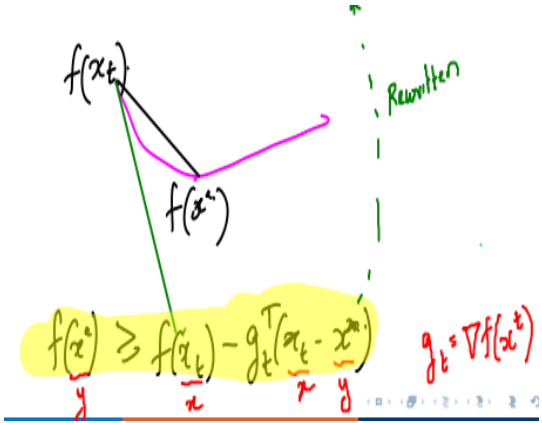
- Let us invoke convexity with $x = x_t, y = x^*$.

$$f(x_t) - f(x^*) \leq g_t^T (x_t - x^*) \tag{2}$$

Part II: Invoking Convexity

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

$$f(x_t) - f(x^*) \leq g_t^T (x_t - x^*) \tag{2}$$



Part II: Invoking Convexity

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

$$f(x_t) - f(x^*) \leq g_t^T (x_t - x^*) \tag{2}$$

- Recall from (1):

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

which, based on $||x_T - x^*||^2 \geq 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T (x_t - x^*) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} ||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \tag{3}$$

Part II: Invoking Convexity

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

$$f(x_t) - f(x^*) \leq g_t^T(x_t - x^*) \tag{2}$$

- Recall from (1):

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

which, based on $||x_T - x^*||^2 \geq 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} ||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \tag{3}$$

Next: Complete the analysis based on all the elements we have already brought together

Part II: Invoking Convexity

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

$$f(x_t) - f(x^*) \leq g_t^T (x_t - x^*) \tag{2}$$

- Recall from (1):

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

  which, based on $||x_T - x^*||^2 \geq 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T (x_t - x^*) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} ||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \tag{3}$$

- Combining (2) with (3), we have:

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

Part II: Invoking Convexity

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

$$f(x_t) - f(x^*) \leq g_t^T (x_t - x^*) \tag{2}$$

- Recall from (1):

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

which, based on $||x_T - x^*||^2 \geq 0$, implies:

$$\sum_{t=0}^{T-1} g_t^T (x_t - x^*) \leq \frac{\gamma}{2} \sum_{t=0}^{T-1} ||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \tag{3}$$

- Combining (2) with (3), we have:

Lipschitz continuity can give us some upper bound on gradient norm

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

Part III: Invoking Lipschitz Continuity

- Recall final result:

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

- Let $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B$ for all x. Then...
- $\frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \leq \frac{\gamma}{2}TB^2 + \frac{R^2}{2\gamma}$
- The extreme right expression $\frac{\gamma}{2}TB^2 + \frac{R^2}{2\gamma}$ is minimized with respect to γ, by setting its derivative (wrt γ) to 0 which is obtained by setting $\gamma = \frac{R}{B\sqrt{T}}$.

Part III: Invoking Lipschitz Continuity

- Recall final result: E_1 (which is independent of r)

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

- Let $||x_0 - x^*|| \leq R$ and $||\nabla f(x)|| \leq B$ for all x. Then...
- $\frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2) \leq \frac{\gamma}{2}TB^2 + \frac{R^2}{2\gamma}$ → $E_2(r)$ to value $\frac{RB}{\sqrt{T}}$.
- The extreme right expression $\frac{\gamma}{2}TB^2 + \frac{R^2}{2\gamma}$ is minimized with respect to γ, by setting its derivative (wrt γ) to 0 which is obtained by setting $\gamma = \frac{R}{B\sqrt{T}}$.

Trick 3: Determine the minimum value of an upper bound (likewise maximum value of lower bound)

E_1 (which is independent of r) ---→ $E_1 \leq E_2(r) \Leftrightarrow E_1 \leq \min E_2(r)$

Part III: Invoking Lipschitz Continuity

- Recall final result:

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

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

$$\frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x^*)] \leq \frac{RB}{\sqrt{T}}$$

- 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.

Part III: Invoking Lipschitz Continuity

- Recall final result:

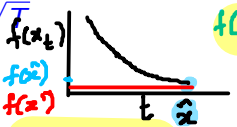
$$\sum_{t=0}^{T-1}(f(x_t) - f(x^*)) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}||g_t||^2 + \frac{1}{2\gamma}(||x_0 - x^*||^2)$$

Note: This inequality holds for any $\gamma > 0$. However, choice of step size can be important for analysis of convergence of some optimization algorithms

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

$$\frac{1}{T}\sum_{t=0}^{T-1}[f(x_t) - f(x^*)] \leq \frac{RB}{\sqrt{T}}$$

1) Recap from Machine Learning that often the iterations are stopped based on validation set (i.e. when validation accuracy starts dropping) Here, the optimization is wrt training set. So we have early stopping for avoiding overfitting/to generalize well
2) Numerical precision can sometimes be a deterrent to get exact equality of $f(x) = f(x^*)$



- 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.

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,

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

Suppose our need is to find T such that

$|f(\hat{x}) - f(x^*)| < \epsilon$

Note that we do not have access to this value!

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

Q: Can this help derive a sufficient condition on T?

Ans: Yes. Just flip numberator and denominator and square both

$\Rightarrow \frac{R^2B^2}{\epsilon^2} \leq T$

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 $|f(\hat{x}) - f(x^*)| \leq \epsilon$, it suffices to have

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

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 $|f(\hat{x}) - f(x^*)| \leq \epsilon$, it suffices to have

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

- Which implies that:

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

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 $|f(\hat{x}) - f(x^*)| \leq \epsilon$, it suffices to have

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

- Which implies that:

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

Optional: Rate&Order of Convergence, Generalized Gradient Descent:

https://moodle.iitb.ac.in/pluginfile.php/143881/mod_resource/content/2/Optional%20Reading%20-%20Q-Convergence%20principles%2Bgeneral%20descent%20algorithms%2Bbacktracking%20ray%20search%2Barmijo%20conditions.pdf

HOMEWORK

$T = O\left(\frac{1}{\epsilon^2}\right)$?

OR

$T = \Omega\left(\frac{1}{\epsilon^2}\right)$?

BIG Omega is about the limiting (lower bound) behaviour of the function for epsilon tending to infinity

This kind of analysis is less relevant for us

See R & Q-convergence

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 $|f(\hat{x}) - f(x^*)| \leq \epsilon$, it suffices to have

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

- Which implies that:

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

- 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 $\frac{R^2B^2}{\epsilon^2}$ iterations.
- Advantages of this bound: a) Goes to zero as T gets large, and b) Independent of the dimensionality of $\mathbf{x}$!

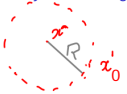
How good or bad is this bound?

- 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) Independent of the dimensionality of $\mathbf{x}$!
- 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 $10^8 R^2 B^2$!

How good or bad is this bound?

The analysis below assumes that we are always dealing with the same initial iterate x_1 (which determines R)

The recipe for number of iterates under this assumption is that if you want to get more close to the optimal, you would need number iterations inversely prportional to the square of how close you need to get to the optimal



Note that R is characterizing the initial iterate's distance only

- 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) Independent of the dimensionality of $\mathbf{x}$! Only the gradient computation will depend on the dimensionality of x
  The analysis is based on unit of each step which is a single gradient computaton
- Disadvantages: Slow convergence. To achieve a an error of 0.01, we require $10^4R^2B^2$ iterations. To achieve an error of 0.0001, the number of iterations is $10^8R^2B^2$!

Other disadvantages of the assumptions underlying this analysis of the algorithm
Does not assume that the step size γ is obtained in a more principled (search based) manner
See extra and optional slides: https://moodle.iitb.ac.in/pluginfile.php/140882/mod_resource/content/2/Optimal%20Rate%20-%20Convergence%20principle%20general%20descent%20algorithm%20backtracking%20ray%20search%20armijo%20conditions.pdf
Realistically gamma can be obtained using search techniques such as exact/backtracking ray search
Specifically backtracking ray search continue until conditions such as Armijo conditions/Goldstein conditions etc are satisfied