Ray Search for Descent: Options

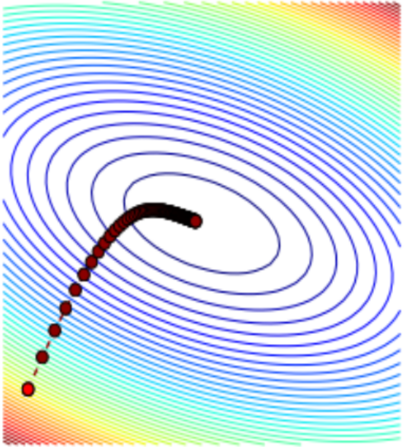
1. **Exact ray search:** The exact ray search seeks a scaling factor t that satisfies

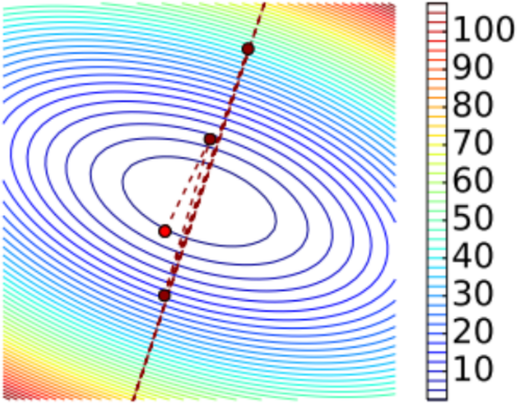
   $$t = \underset{t>0}{\operatorname{argmin}} f(\mathbf{x} + t\Delta\mathbf{x}) \tag{28}$$

2. **Backtracking ray search:** The exact line search may not be feasible or could be expensive to compute for complex non-linear functions. A relatively simpler ray search iterates over values of step size starting from 1 and scaling it down by a factor of $\beta \in (0, \frac{1}{2})$ after every iteration till the following condition, called the *Armijo condition* is satisfied for some $0 < c_1 < 1$.

   $$f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) + c_1 t \nabla^T f(\mathbf{x}) \Delta\mathbf{x} \tag{29}$$

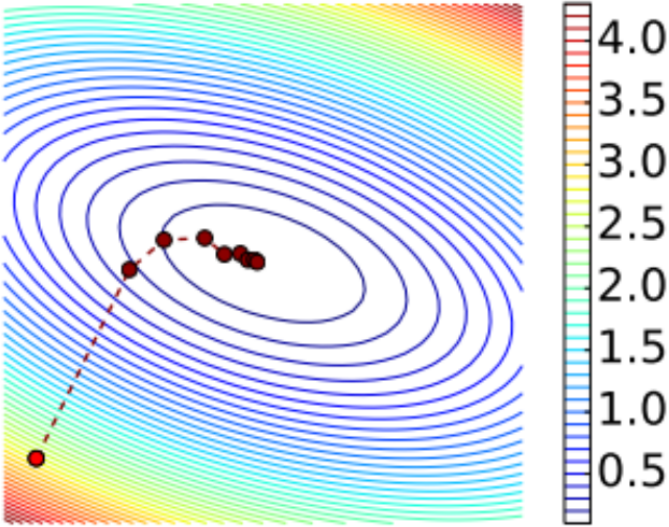
   Based on first order convexity condition, it can be inferred that when $c_1 = 1$, the right hand side of (29) gives a lower bound on the value of $f(\mathbf{x} + t\Delta\mathbf{x})$ and hence





Oscillations in f(x)
as well as in x

Change in f(x) somewhat fast
but could be faster
Small change in x
Seems to implement Armijo condition



Starts fast and then adaptively slows down

Hope for this behaviour with the combination of the two conditions

Ray Search for First Order Descent: Strong Wolfe Conditions

Wolfe's condition: The function should have a **sufficient rate of decrease**.

third plot seemed to be benefiting from this

$$\left|\Delta \mathbf{x}^T \nabla f(\mathbf{x} + t\Delta \mathbf{x})\right| \leq c_2 \left|\Delta \mathbf{x}^T \nabla f(\mathbf{x})\right| \tag{30}$$

where $1 > c_2 > c_1 > 0$. This condition ensures that the slope of the function $f(\mathbf{x} + t\Delta\mathbf{x})$ at t is less than c_2 times that at $t = 0$.

1. The conditions in **(29)** and **(30)** are together called the strong Wolfe conditions. These conditions are particularly very important for non-convex problems.
2. While **(29) ensures guaranteed decrease in** $f(\mathbf{x} + \Delta\mathbf{x})$ in terms of the slope, **(30) provides guaranteed decrease in magnitude of slope and (indirectly) avoids too small steps**.
3. Claim: If $1 > c_2 > c_1 > 0$ and the function $f(\mathbf{x})$ is convex and differentiable, there exists t such that (29) and (30) are both satisfied for any f. *Hint: Use the Mean Value Theorem*

Algorithms: Steepest Descent

- The idea of steepest descent is to determine a descent direction such that for a unit step in that direction, the prediction of decrease in the objective is maximized
- However, consider $\Delta x = \text{argmin}_v \begin{bmatrix} -5 & 10 & 15 \end{bmatrix} v$

  $$\implies \Delta x = \begin{bmatrix} \infty \\ -\infty \\ -\infty \end{bmatrix}$$

  which is unacceptable
- Thus, there is a necessity to restrict the norm of v
- The choice of the descent direction can be stated as:

$$\Delta x = \underset{v}{\text{argmin}}\, \nabla^\top f(x) v$$

s.t. $\|v\| = 1$

Various choices of the norm result in different solutions for Δx

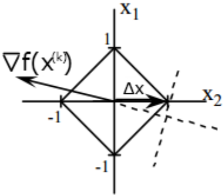
- For 2-norm, $\Delta x = -\frac{\nabla f(x^{(k)})}{\left\|\nabla f(x^{(k)})\right\|_2}$

  (gradient descent)

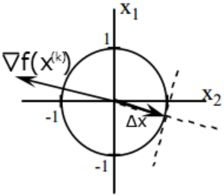
- For 1-norm, $\Delta x = -\operatorname{sign}\left(\frac{\partial f(x^{(k)})}{\partial x_i^{(k)}}\right) e_i$, where e_i is the i^{th} standard basis vector and i is the component $\frac{\partial f(x^{(k)})}{\partial x_i^{(k)}}$ with the maximum magnitude

  (coordinate descent)

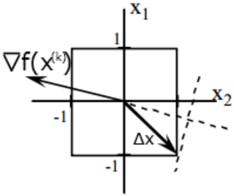
- For ∞-norm, $\Delta x = -\operatorname{sign}(\nabla f(x^{(k)}))$



SDD in L1-norm



SDD in L2-norm



SDD in L∞ -norm

For $\mathbf{x} = (x_1, x_2)^T \in \mathbb{R}^2$, the *Rosenbrock function* is defined as

$$f(\mathbf{x}) := 100(x_2 - x_1^2)^2 + (1 - x_1)^2.$$

Show that the only solution of $\nabla f(\mathbf{x}) = \mathbf{0}$ is $\mathbf{x}^* = (1,1)^T$ (that is, $x_1 = x_2 = 1$);

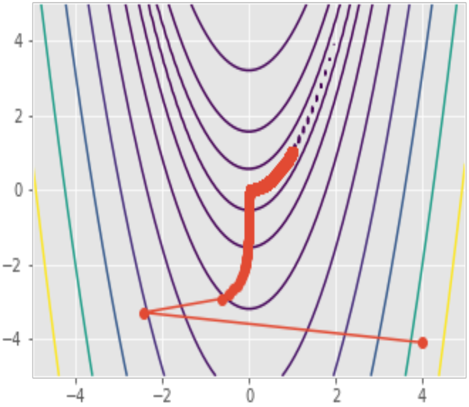
Is the function $f(\mathbf{x})$ convex on $\mathbb{R}^2$? Explain. (No - see function plots and sublevel sets in the following slides)

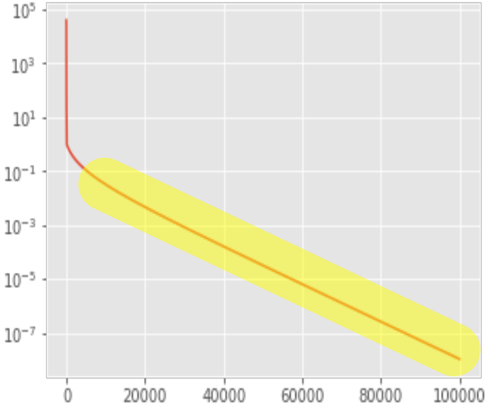
Generalizing the Rosenbrock function to n dimensions

$$f(x_1 \cdots x_n) = \sum_{i=1}^{n-1} (100(x_i^2 - x_{i+1})^2 + (1 - x_i)^2)$$

minimum at $f(1, 1, \cdots, 1) = 0$

Gradient descent on the Rosenbrock function converges but with a poor rate





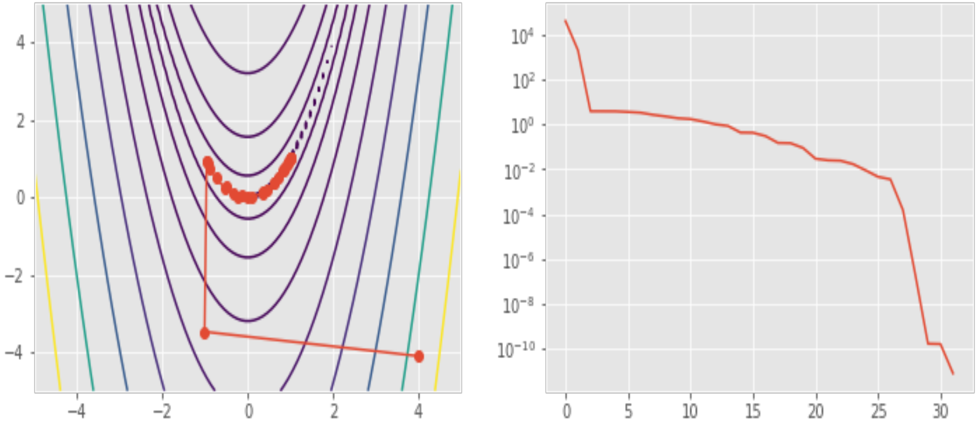
after a point, linear convergence (unacceptable)

Gradient descent on the Rosenbrock function converges but with a poor rate

```
fun: 1.0604663473448339e-08
nfev: 100001
nit: 100000
 success: True
x: array([ 0.9999,  0.9998])
```

Too slow!

Conjugate gradient (gradient adapted according to a quadratic estimation of the curvature where the quadratic estimation it itself getting adapted)



Optional: For understanding CG, see Figure 4.55 (and pages 317-323) of https://www.cse.iitb.ac.in/~cs709/notes/BasicsOfConvexOptimization.pdf

Conjugate gradient (gradient adapted according to a quadratic estimation of the curvature where the quadratic estimation it itself getting adapted)

```
     fun: 7.976921523473763e-12
     jac: array([ -9.4059e-07,  -2.3516e-06])
 message: 'Optimization terminated successfully.'
    nfev: 70
     nit: 31
    njev: 70
  status: 0
 success: True
       x: array([ 1.,  1.])
```

Much faster

General Algorithm: Steepest Descent (contd)

```
**Find** a starting point $\mathbf{x}^{(0)} \in \mathcal{D}$.
**repeat**
    1. Set $\Delta\mathbf{x}^{(k)} = \text{argmin}\left\{\nabla^T f(\mathbf{x}^{(k)})\mathbf{v} \mid ||\mathbf{v}|| = 1\right\}$.
    2. Choose a step size $t^{(k)} > 0$ using exact or backtracking ray search.
    3. Obtain $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t^{(k)}\Delta\mathbf{x}^{(k)}$.
    4. Set $k = k + 1$.
**until** stopping criterion (such as $||\nabla f(\mathbf{x}^{(k+1)})|| \leq \epsilon$) is satisfied
```

As we have seen, choice of descent direction can make a big difference!

Figure 8: The steepest descent algorithm.

Two examples of the steepest descent method are the gradient descent method (for the eucledian or L_2 norm) and the coordinate-descent method (for the L_1 norm). One fact however is that no two norms should give exactly opposite steepest descent directions, though they may point in different directions.

Convergence of Descent Algorithm

- Consider the general descent algorithm ($\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k) with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k \Delta\mathbf{x}^k$.
  - Suppose f is bounded below in $\Re^n$ and
  - is continuously differentiable in an open set $\mathcal{N}$ containing the level set $\{\mathbf{x} | f(\mathbf{x}) \le f(\mathbf{x}^0)\}$
  - ∇f is Lipschitz continuous.

  Then, $\displaystyle\sum_{k=1}^{\infty} \frac{(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k)^2}{\|\Delta\mathbf{x}^k\|^2} < \infty$ (that is, it is finite)
- Thus, $\lim_{k\to\infty} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|} = 0$.
- If we additionally assume that the descent direction is not orthogonal to the gradient, *i.e.*, $-\frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|\|\nabla f(\mathbf{x}^k)\|} \ge \Gamma$ for some $\Gamma > 0$, then, we can show that $\lim_{k\to\infty} \|\nabla f(\mathbf{x}^k)\| = 0$
- Before we try and prove this result, let us discuss Lipschitz continuity (recall from midsem).

Lipschitz Continuity

Recall: Lipschitz Continuity of f

- Formally, $f(x) : \mathcal{D} \subseteq \Re^n \rightarrow \Re$ is Lipschitz continuous if $|f(\mathbf{x}) - f(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{D}$.
- A Lipschitz continuous function is limited in how fast it changes: there exists a definite positive real number $L > 0$ such that, for every pair of points on the graph of the function, the absolute value of the slope of the line connecting them is not greater than this real number. This bound is called the function's Lipschitz constant, $L > 0$.
- We can show that if a function $f : \Re \rightarrow \Re$ is convex in (α, β) it is Lipschitz continuous in $[\gamma, \delta]$ where $\alpha < \gamma < \delta < \beta$. We do not assume that f is differentiable.

Lipschitz continuous for a fixed alpha, beta

Convex Function is Lipschitz continuous

$f(x) : \Re \to \Re$ is Lipschitz continuous in $[\gamma, \delta]$ if $|f(x) - f(y)| \le L|x - y|$ for all $x, y \in [\gamma, \delta]$. We will show that if a function $f : \Re \to \Re$ is convex in (α, β) it is Lipschitz continuous in $[\gamma, \delta]$ where $\alpha < \gamma < \delta < \beta$. Do not assume that f is differentiable. Fill up the three blanks below.

- Let $p, q \in \Re$ such that $\alpha < p < \gamma < \delta < q < \beta$ and let $x_1, x_2 \in [\gamma, \delta]$. Then

  $$\underline{\hspace{8cm}}_1 \le \frac{f(x_2) - f(x_1)}{x_2 - x_1} \le \underline{\hspace{7cm}}$$

  because of convexity of f.
- Take $L = \underline{\hspace{8cm}}_3$

to prove Lipschitz continuity of f in the interval $[\gamma, \delta]$.

Convex Function is Lipschitz continuous (Locally)

$f(x) : \Re \to \Re$ is Lipschitz continuous in $[\gamma, \delta]$ if $|f(x) - f(y)| \le L|x - y|$ for all $x, y \in [\gamma, \delta]$. We will show that if a function $f : \Re \to \Re$ is convex in (α, β) it is Lipschitz continuous in $[\gamma, \delta]$ where $\alpha < \gamma < \delta < \beta$. Do not assume that f is differentiable. Fill up the three blanks below.

- Let $p, q \in \Re$ such that $\alpha < p < \gamma < \delta < q < \beta$ and let $x_1, x_2 \in [\gamma, \delta]$. Then

$$\underline{\hspace{9cm}}_1 \le \frac{f(x_2) - f(x_1)}{x_2 - x_1} \le \underline{\hspace{8cm}}$$

  because of convexity of f.

- Take $L = \underline{\hspace{9cm}}_3$

to prove Lipschitz continuity of f in the interval $[\gamma, \delta]$.

1. $\underline{\hspace{9cm}}_1 = \frac{f(\gamma)-f(p)}{\gamma-p}$
2. $\underline{\hspace{9cm}}_2 = \frac{f(q)-f(\delta)}{q-\delta}$
3. $\underline{\hspace{9cm}}_3 = \max\left\{\left|\frac{f(\gamma)-f(p)}{\gamma-p}\right|, \left|\frac{f(q)-f(\delta)}{q-\delta}\right|\right\}$

This L is dependent on p, q , gamma, delta (that is on alpha, beta)

L is relative to the interval Local Lipschitz continuity

Lipschitz continuity

- Intuitively, a Lipschitz continuous function is limited in how fast it changes: there exists a definite real number L such that, for every pair of points on the graph of the gradient, the absolute value of the slope of the line connecting them is not greater than this real number
  - This bound is called the function's Lipschitz constant, $L > 0$
  - The sum of two Lipschitz continuous functions is also Lipschitz continuous with the Lipschitz constant specified as the sum of the respective Lipschitz constants.

Lipschitz continuity

- Intuitively, a Lipschitz continuous function is limited in how fast it changes: there exists a definite real number L such that, for every pair of points on the graph of the gradient, the absolute value of the slope of the line connecting them is not greater than this real number
  - This bound is called the function's Lipschitz constant, $L > 0$
  - The sum of two Lipschitz continuous functions is also Lipschitz continuous with the Lipschitz constant specified as the sum of the respective Lipschitz constants.
  - The product of two Lipschitz continuous and bounded functions is also Lipschitz continuous
- Now, $\nabla f(x)$ is Lipschitz continuous if $\left\|\nabla f(x) - \nabla f(y)\right\| \leq L\|x - y\|$

f1(x)=f2(x)=x
are Lipschitz
continuous
But not
f1*f2 = x^2

Recap how we generalized monotonicity
from scalar valued to vector valued functions
Something similar here...

Interpretation of Lipschitz continuity of $\nabla f(\mathbf{x})$

- Consider $\nabla f(x) \in \mathbf{R}$, and $\nabla f(x) = \frac{df}{dx} = f'(x)$
- $|f'(x) - f'(y)| \leq L|x - y|$

Interpretation of Lipschitz continuity of $\nabla f(\mathbf{x})$

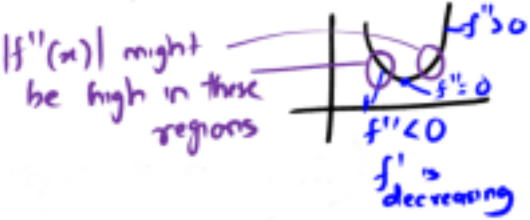
- Consider $\nabla f(x) \in \mathbf{R}$, and $\nabla f(x) = \frac{df}{dx} = f'(x)$
- $|f'(x) - f'(y)| \leq L|x - y|$

  $$\implies \frac{f'(x)-f'(y)}{|x-y|} \leq L$$

  $$\implies \left|\frac{f'(x+h)-f'(x)}{h}\right| \leq L \quad \text{(putting } \underline{y = x + h}\text{)}$$
- Taking limit $\underline{h \to 0}$, we get

Interpretation of Lipschitz continuity of $\nabla f(\mathbf{x})$

- Consider $\nabla f(x) \in \mathbf{R}$, and $\nabla f(x) = \frac{df}{dx} = f'(x)$
- $|f'(x) - f'(y)| \leq L|x-y|$
  $\implies \frac{f'(x)-f'(y)}{|x-y|} \leq L$
  $\implies \left|\frac{f'(x+h)-f'(x)}{h}\right| \leq L \quad$ (putting $y = x + h$)
- Taking limit $h \to 0$, we get
  $\underline{|f''(x)| \leq L}$ (assuming the limit exits)
- f'' represents curvature



Recap: Strong convexity was about guatanteed lower bounded curvature

Lipschitz Continuity of $\nabla f(\mathbf{x})$ and Hessian

- Let $f(\mathbf{x})$ have continuous partial derivatives and continuous mixed partial derivatives in an open ball $\mathcal{R}$ containing a point $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$.
- Let $\nabla^2 f(\mathbf{x})$ denote an $n \times n$ matrix of mixed partial derivatives of f evaluated at the point $\mathbf{x}$, such that the ij^{th} entry of the matrix is $f_{x_i x_j}$. The matrix $\nabla^2 f(\mathbf{x})$ is called the Hessian matrix.
- The Hessian matrix is symmetric[4].

[4] By Clairaut's Theorem, if the partial and mixed derivatives of a function are continuous on an open region containing a point $\mathbf{x}^*$, then $f_{x_i x_j}(\mathbf{x}^*) = f_{x_j x_i}(\mathbf{x}^*)$, for all $i, j \in [1, n]$.

Lipschitz Continuity of $\nabla f(\mathbf{x})$ and Hessian

- Let $f(\mathbf{x})$ have continuous partial derivatives and continuous mixed partial derivatives in an open ball $\mathcal{R}$ containing a point $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$.
- Let $\nabla^2 f(\mathbf{x})$ denote an $n \times n$ matrix of mixed partial derivatives of f evaluated at the point $\mathbf{x}$, such that the ij^{th} entry of the matrix is $f_{x_i x_j}$. The matrix $\nabla^2 f(\mathbf{x})$ is called the Hessian matrix.
- The Hessian matrix is symmetric[4].
- For a Lipschitz continuous $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$, we can show that for any vector v,

[4] By Clairaut's Theorem, if the partial and mixed derivatives of a function are continuous on an open region containing a point $\mathbf{x}^*$, then $f_{x_i x_j}(\mathbf{x}^*) = f_{x_j x_i}(\mathbf{x}^*)$, for all $i, j \in [1, n]$.

Lipschitz Continuity of $\nabla f(\mathbf{x})$ and Hessian

- Let $f(\mathbf{x})$ have continuous partial derivatives and continuous mixed partial derivatives in an open ball $\mathcal{R}$ containing a point $\mathbf{x}^*$ where $\nabla f(\mathbf{x}^*) = 0$.
- Let $\nabla^2 f(\mathbf{x})$ denote an $n \times n$ matrix of mixed partial derivatives of f evaluated at the point $\mathbf{x}$, such that the ij^{th} entry of the matrix is $f_{x_i x_j}$. The matrix $\nabla^2 f(\mathbf{x})$ is called the Hessian matrix.
- The Hessian matrix is symmetric[4].
- For a Lipschitz continuous $\nabla f : \mathbf{R}^n \to \mathbf{R}^n$, we can show that for any vector v,
  - $v^\top \nabla^2 f(x) v \leq v^\top L v$
    $\implies v^\top (\nabla^2 f(x) - LI) v \leq 0$
  - That is, $\nabla^2 f(x) - LI$ is negative semi-definite
  - This can be written as:

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

[4] By Clairaut's Theorem, if the partial and mixed derivatives of a function are continuous on an open region containing a point $\mathbf{x}^*$, then $f_{x_i x_j}(\mathbf{x}^*) = f_{x_j x_i}(\mathbf{x}^*)$, for all $i, j \in [1, n]$.

Example: $f(x) = \frac{x^3}{3}$

- $f(x) = \frac{x^3}{3} \implies f'(x) = x^2$
- **Claim:** $f'(x)$ is locally Lipschitz continuous but not globally

f''(x) = 2x not upper bounded globally
But
In a fixed interval (x-1,x+1), upper bounded by 2|x|

Example: $f(x) = \frac{x^3}{3}$

- $f(x) = \frac{x^3}{3} \implies f'(x) = x^2$
- **Claim:** $f'(x)$ is locally Lipschitz continuous but not globally
- Consider $x \in \mathbf{R}$
- $\sup_{y\in(x-1,x+1)} |f''(y)| = \sup_{y\in(x-1,x+1)} |2y| \leq 2|x| + 1$
- Applying mean value theorem for $(y, z) \in (x - 1, x + 1)$:

Example: $f(x) = \frac{x^3}{3}$

- $f(x) = \frac{x^3}{3} \implies f'(x) = x^2$
- **Claim:** $f'(x)$ is locally Lipschitz continuous but not globally
- Consider $x \in \mathbf{R}$
- $\sup_{y\in(x-1,x+1)} |f''(y)| = \sup_{y\in(x-1,x+1)} |2y| \leq 2|x| + 1$
- Applying mean value theorem for $(y, z) \in (x - 1, x + 1)$: $\exists\, \lambda$ such that $f''(\lambda) = \frac{f'(y)-f'(z)}{y-z}$

- $|f'(y) - f'(z)| = |f''(\lambda)(y - z)|$
  $\leq \big|2|x| + 1\big|\,|y - x|$, $\forall (y, z) \in (x - 1, x + 1)^2$
- Thus, $L = \big|2|x| + 1\big|$
- Therefore, f is locally Lipschitz continuous

- $|f'(y) - f'(z)| = |f''(\lambda)(y-z)|$
  $\leq \big|2|x|+1\big|\,|y-x|,\ \forall (y,z) \in (x-1, x+1)^2$
- Thus, $L = \big|2|x|+1\big|$
- Therefore, f' is Lipschitz continuous in $(x-1, x+1)$
- But as $x \to \infty$, $L \to \infty$
- This implies that f' may not be Lipschitz continuous everywhere
- Consider $y \neq 0$, and
  $\frac{f'(y)-f'(0)}{|y-0|} = |y|$
- $\underline{|y| \to \infty}$ as $y \to \infty$
- Thus, f' is proved to not be Lipschitz continuous globally

Lipschitz Continuity: Another example

- Consider

Show (H/W)

$$f(x) = \begin{cases} x^2 sin\left(\frac{1}{x^2}\right) & \text{if } x \neq 0 \\ 0 & \text{if } x = 0 \end{cases}$$

- We can verify that this function is continuous and differentiable everywhere *i.e.* $f''(0) = 0$ from left and right
- However, we can show that $f(x)$ is *not* Lipschitz continuous

Lipschitz continuity: Another example

- **Consider:** $f'(x) = |x|$
- Since $|f'(x) - f'(y)| = \big||x| - |y|\big| \leq |x - y|$,
  f' is Lipschitz continuous with $L = 1$
- However, it is not differentiable everywhere (not at 0)
- In fact, if f is continuously differentiable everywhere, it is also Lipschitz continuous
- For functions over a closed and bounded subset of the real line: f is continuous $\supseteq$ f is differentiable (almost everywhere) $\supseteq$ f is Lipschitz continuous $\supseteq$ f' is continuous $\supseteq$ f' is differentiable
- Recap (now generalized to $f : \Re^n \to \Re$) that f is locally Lipschitz continuous $\supseteq$ f is convex

←

IMPLIES

Considering gradients in Lipschitz continuity

- If ∇f is Lipschitz continuous, then

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

- **Taylor's theorem** states that if f and its first n derivatives $f', f'', \ldots, f^{(n)}$ are continuous in the closed interval $[a, b]$, and differentiable in (a, b), then there exists a number $c \in (a, b)$ such that

$$f(b) = f(a) + f'(a)(b-a) + \frac{1}{2!}f''(a)(b-a)^2 + \ldots + \frac{1}{n!}f^{(n)}(a)(b-a)^n + \frac{1}{(n+1)!}f^{(n+1)}(c)(b-a)^{n+1}$$

Approximation if you ignore last term
Last term is in terms of a c in (a,b)

- We will invoke Taylor's theorem up to the second degree:

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2}f''(c)(y-x)^2$$

  where $c \in (x, y)$ and $x, y \in \mathbf{R}$

- Let us generalize to $f : \mathbf{R}^n \to \mathbf{R}$:

- We will invoke Taylor's theorem up to the second degree:

$$f(y) = f(x) + f'(x)(y-x) + \frac{1}{2}f''(c)(y-x)^2$$

where $c \in (x, y)$ and $x, y \in \mathbf{R}$

All this comes from the basic mean value theorem

- Let us generalize to $f : \mathbf{R}^n \to \mathbf{R}$:

$$\underline{f(\mathbf{y}) = f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{1}{2}(\mathbf{y}-\mathbf{x})^T\nabla^2 f(\mathbf{c})(\mathbf{y}-\mathbf{x})}$$

where $\underline{\mathbf{c} = \mathbf{x} + \Gamma(\mathbf{y}-\mathbf{x})}$, $\Gamma \in (0,1)$, and $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$

- If ∇f is Lipschitz continuous and f is doubly differentiable,

Contrast with strong convexity?

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \qquad (36)$$

This inequality can be shown to be another condition for Lipschitz continuity (without requiring double differentiability)

- We will invoke Taylor's theorem up to the second degree:

$$f(y) = f(x) + f'(x)(y - x) + \frac{1}{2}f''(c)(y - x)^2$$

  where $c \in (x, y)$ and $x, y \in \mathbf{R}$
- Let us generalize to $f : \mathbf{R}^n \to \mathbf{R}$:

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

  where $\mathbf{c} = \mathbf{x} + \Gamma(\mathbf{y} - \mathbf{x})$, $\Gamma \in (0, 1)$, and $\mathbf{x}, \mathbf{y} \in \mathbf{R}^n$
- If ∇f is Lipschitz continuous and f is doubly differentiable,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{y} - \mathbf{x}\|^2 \tag{36}$$

- While we showed (36) assuming f is doubly differentiable, (36) holds for any Lipschitz continuous $\nabla f(\mathbf{x})$.

Gradient Descent and Lipschitz Continuity

1. Replacing $\mathbf{x}$ by $\mathbf{x}^k$ and y by the gradient descent update $\mathbf{x}^{k+1} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k)$, and applying necessary condition for Lipschitz continuity:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}) + \frac{L}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2$$

2. For a descent algorithm, $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k = \nabla^T f(\mathbf{x}^k)\Delta(\mathbf{x}^{k+1} - \mathbf{x}^k) < 0$ for each k
3. Putting together steps 1 and 2 above,

Gradient Descent and Lipschitz Continuity

1. Replacing $\mathbf{x}$ by $\mathbf{x}^k$ and y by the gradient descent update $\mathbf{x}^{k+1} = \mathbf{x}^k - t\nabla f(\mathbf{x}^k)$, and applying necessary condition for Lipschitz continuity:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}) + \frac{L}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2$$

2. For a descent algorithm, $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k = \nabla^T f(\mathbf{x}^k)\Delta(\mathbf{x}^{k+1} - \mathbf{x}^k) < 0$ for each k
3. Putting together steps 1 and 2 above,

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \frac{L}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 \qquad (37)$$