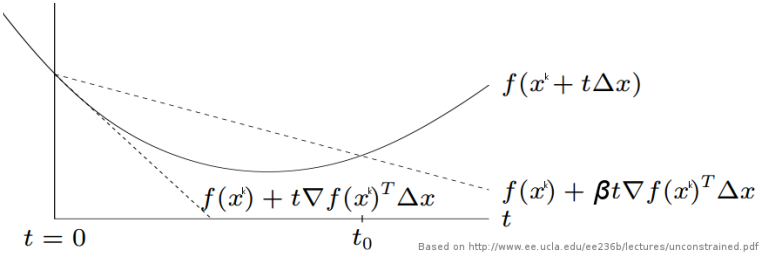
Interpretation of backtracking line search [RECAP]



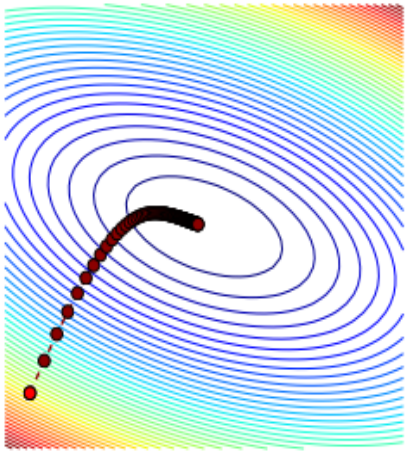
Based on http://www.ee.ucla.edu/ee236b/lectures/unconstrained.pdf

- Δx = direction of descent = $-\nabla f(x^k)$ for gradient descent
- A different way of understanding the varying step size with β: Multiplying t by β causes the interpolation to tilt as indicated in the figure

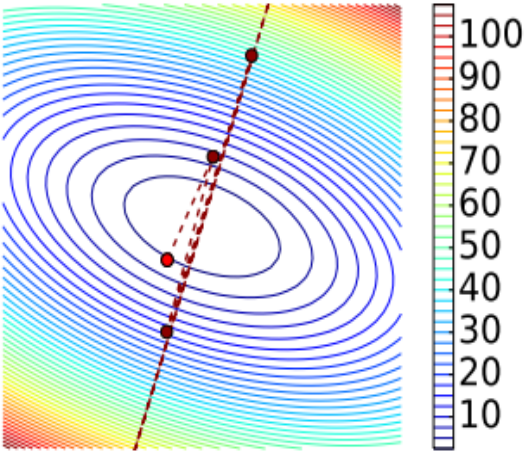
Homeworks: Let $f(x) = x^2$ for $x \in \Re$. Let $x^0 = 2$, $\Delta x^k = -1$ for all k (since it is a valid descent direction of $x > 0$) and $x^k = 1 + 2^{-k}$. What is the step size t^k implicitly being used. Show that while t^k satisifies the Armijo condition (determine a c_1), it does not satisfy the second Strong Wolfe condition in the following slides. Why is the choice of step size bad?

Strong Wolfe conditions: Sufficient decrease BUT NOT at the cost of oscillations

Small step size



Large step size



Ray Search for First Order Descent: Strong Wolfe Conditions [RECAP]

Often, another condition is used for inexact line search in conjunction with the Armijo condition.

$$\left|\Delta \mathbf{x}^T \nabla f(\mathbf{x} + t\Delta \mathbf{x})\right| \leq c_2 \left|\Delta \mathbf{x}^T \nabla f(\mathbf{x})\right| \tag{28}$$

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 (27) and (28) are together called the strong Wolfe conditions. These conditions are particularly very important for non-convex problems.
2. While (**??**refeq:armijoCondition) ensures guaranteed decrease in $f(\mathbf{x} + \delta \mathbf{x})$, (28) provides guaranteed decrease in magnitude of slope and avoid 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 (27) and (28) are both satisfied for any f. *Hint: Use the Mean Value Theorem*

Convexity $\Rightarrow$ Strong Wolfe Conditions [RECAP]

- Let $\phi(t) = f(\mathbf{x}^k + t\Delta\mathbf{x}^k) \geq f(\mathbf{x}^k) + t\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ (where the second inequality is by virtue of convexity). Remember that $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$
- Since $0 < c_1 < 1$, the linear approximation $l(t) = f(\mathbf{x}^k) + tc_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ is unbounded below and it can be shown to

Convexity $\Rightarrow$ Strong Wolfe Conditions [RECAP]

- Let $\phi(t) = f(\mathbf{x}^k + t\Delta\mathbf{x}^k) \geq f(\mathbf{x}^k) + t\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ (where the second inequality is by virtue of convexity). Remember that $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$
- Since $0 < c_1 < 1$, the linear approximation $l(t) = f(\mathbf{x}^k) + tc_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ is unbounded below and it can be shown to intersect the graph of ϕ atleast once.
- Let $t' > 0$ be the smallest intersecting value of t, that is:

$$f(\mathbf{x} + t'\Delta\mathbf{x}^k) = f(\mathbf{x}^k) + t'c_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{29}$$

- For all $t \in [0, t']$,

Convexity $\Rightarrow$ Strong Wolfe Conditions [RECAP]

- Let $\phi(t) = f(\mathbf{x}^k + t\Delta\mathbf{x}^k) \geq f(\mathbf{x}^k) + t\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ (where the second inequality is by virtue of convexity). Remember that $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$
- Since $0 < c_1 < 1$, the linear approximation $l(t) = f(\mathbf{x}^k) + tc_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ is unbounded below and it can be shown to intersect the graph of ϕ atleast once.
- Let $t' > 0$ be the smallest intersecting value of t, that is:

$$f(\mathbf{x} + t'\Delta\mathbf{x}^k) = f(\mathbf{x}^k) + t'c_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{29}$$

- For all $t \in [0, t']$,

$$f(\mathbf{x}^k + t\Delta\mathbf{x}^k) \leq f(\mathbf{x}^k) + tc_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{30}$$

  That is, there exists a non-empty set of t such that the first Wolfe condition is met.
- By the mean value theorem, $\exists\, t'' \in (0, t')$ such that

$$f(\mathbf{x}^k + t'\Delta\mathbf{x}^k) - f(\mathbf{x}^k) = t'\nabla^T f(\mathbf{x}^k + t''\Delta\mathbf{x}^k)\Delta\mathbf{x}^k \tag{31}$$

Convexity $\Rightarrow$ Strong Wolfe Conditions (contd.)

- Combining (29) and (31), and using $c_1 < c_2$,and $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$

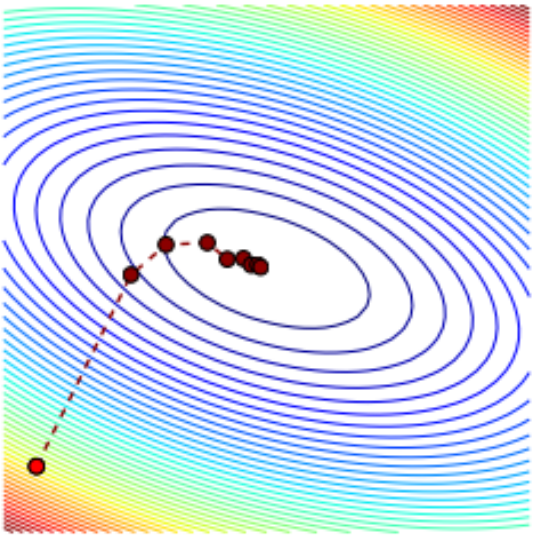
$$\nabla^T f(\mathbf{x}^k + t''\Delta\mathbf{x}^k)\Delta\mathbf{x}^k = c_1\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k > c_2\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{32}$$

- Again, since $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$, we get the $t^k = t''$ satisfying (28)

$$|\nabla^T f(\mathbf{x}^k + t''\Delta\mathbf{x}^k)\Delta\mathbf{x}^k| < c_2|\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k| \tag{33}$$

- In fact, by continuity of $f(.)$, there exists an interval around t'' for which Strong Wolfe conditions hold.

This is desirable: Neither too slow, nor too wayward..



Starts fast and then adaptively slows down

Empirical Observations on Ray Search [RECAP]

- A finding that is borne out of plenty of empirical evidence is that exact ray search does better than empirical ray search in a few cases only.
- Further, the exact choice of the value of c_1 and c_2 seems to have little effect on the convergence of the overall descent method.
- The trend of specific descent methods has been like a parabola - starting with simple steepest descent techniques, then accomodating the curvature hessian matrix through a more sophisticated Newton's method and finally, trying to simplify the Newton's method through approximations to the hessian inverse, culminating in conjugate gradient techniques, that do away with any curvature matrix whatsoever, and form the internal combustion engine of many sophisticated optimization techniques today.
- We start the thread by describing the steepest descent methods.

Algorithms: Steepest Descent [RECAP]

- 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. $\|v\| = 1$

Algorithms: Steepest Descent [RECAP]

- Let $\mathbf{v} \in \Re^n$ be a unit vector under some norm. By first order convexity condition for convex and differentiable f,

$$f(\mathbf{x}^{(k)}) - f(\mathbf{x}^{(k)} + \mathbf{v}) \leq -\nabla^T f(\mathbf{x}^{(k)})\mathbf{v}$$

- For small $\mathbf{v}$, the inequality turns into approximate equality. The term $-\nabla^T f(\mathbf{x}^{(k)})\mathbf{v}$ can be thought of as (an upper-bound on) the first order prediction of decrease.
- The idea in the steepest descent method is to choose a norm and then determine a descent direction such that for a unit step in that norm, the first order prediction of decrease is maximized. This choice of the descent direction can be stated as

$$\Delta\mathbf{x} = \text{argmin}\left\{\nabla^T f(\mathbf{x})\mathbf{v} \mid ||\mathbf{v}|| = 1\right\}$$

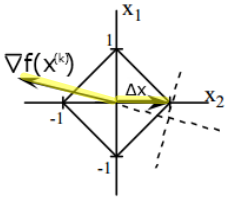
- *Empirical observation:* If the norm chosen is aligned with the gross geometry of the sub-level sets[3], the steepest descent method converges faster to the optimal solution. Else, it often amplifies the effect of oscillations.

[3] The alignment can be determined by fitting, for instance, a quadratic to a sample of the points.

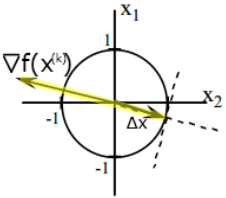
Various choices of the norm result in different solutions for Δx [RECAP]

- For 2-norm, $\Delta x = -\frac{\nabla f(x^{(k)})}{\left\|\nabla f(x^{(k)})\right\|_2}$
  (gradient descent)
- For 1-norm, $\Delta x = -\text{sign}\left(\frac{\partial f(x^{(k)})}{\partial x_i^{(k)}}\right) e_i$, where e_i is the i*th* standard basis vector
  (coordinate descent)
- For ∞-norm, $\Delta x = -\text{sign}(\nabla f(x^{(k)}))$

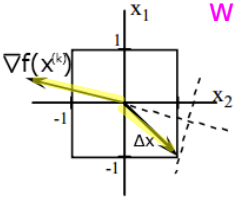
You dont expect the steepest descent direction to have an acute angle with the gradient



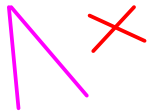
SDD in L1-norm



SDD in L2-norm



SDD in L∞ -norm



General Algorithm: Steepest Descent (contd) [RECAP]

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

Figure 8: The steepest descent algorithm.

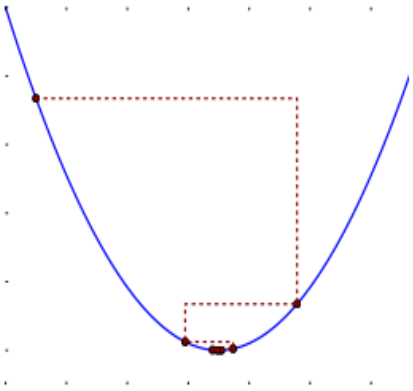
Two examples of the steepest descent method are the gradient descent method (for the eucledian or L_2 norm) and the coordinate-descent method (for the L_1 norm). One fact however is that no two norms should give exactly opposite steepest descent directions, though they may point in different directions.

Convergence of Descent Algorithm To be revisited

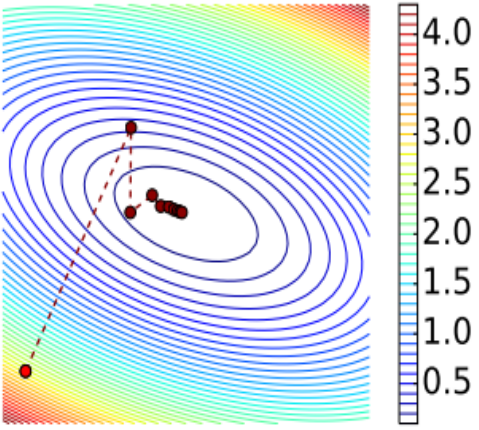
- Consider the general descent algorithm ($\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k) with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k\Delta\mathbf{x}^k$.
  - Suppose f is bounded below in $\Re^n$ and
  - is continuously differentiable in an open set $\mathcal{N}$ containing the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

  Then, $\sum_{k=1}^{\infty} \frac{(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k)^2}{\|\Delta\mathbf{x}^k\|^2} < \infty$ (that is, it is finite) An infinite sum will be 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.*, $-\frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|\|\nabla f(\mathbf{x}^k)\|} \geq \Gamma$ for some $\Gamma > 0$, then, we can show that $\lim_{k\to 0} \|\nabla f(\mathbf{x}^k)\| = 0$
- Before we try and prove this result, let us discuss Lipschitz continuity (recall from midsem).

$f : \mathbb{R} \to \mathbb{R}$



$f : \mathbb{R}^2 \to \mathbb{R}$



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 showed 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 did not assume that f is differentiable.

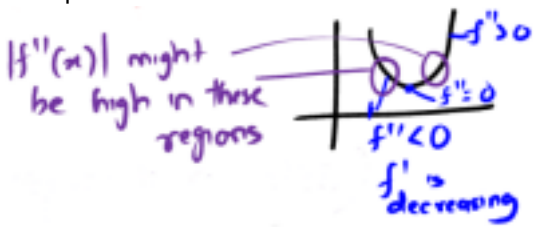
Recap from midsem...

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

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

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



Providing intuition through necessary condition only..

If curvature is high, chances of descent algo oscillating are also high... hence convergence may take time if L is high

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]. If the second derivatives are continuous the Hessian will be symmetric..

[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 is in general unbounded above...
Hence f'(x) cannot be globally Lipschitz...

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:
  $\exists\, (y, z) \in (x-1, x+1)^2, \lambda$
  $f''(\lambda) = \frac{f'(y) - f'(z)}{y - z}$

- $|f'(y) - f'(z)| = |f''(\lambda)(y - z)|$
  $\leq \left|2|x| + 1\right| |y - x|, \forall (y, z) \in (x - 1, x + 1)^2$
- Thus, $L = \left|2|x| + 1\right|$
- Therefore,

- $|f'(y) - f'(z)| = |f''(\lambda)(y - z)|$
  $\leq \left|2|x| + 1\right| |y - x|, \forall (y, z) \in (x - 1, x + 1)^2$
- Thus, $L = \left|2|x| + 1\right|$
- 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
  $\frac{f'(y) - f'(0)}{|y - 0|} = |y|$
- $|y| \to \infty$ as $y \to \infty$
- Thus, f' is proved to not be Lipschitz continuous globally

Another example

- Consider

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

- 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 from midsem (generalized now to $f : \Re^n \to \Re$) that f is Lipschitz continuous $\supseteq$ f is convex

Considering gradients in Lipschitz continuity

Why is the curvature in terms of Hessian upper bounded for Lipschitz continuous functions?

- 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

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

- 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,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \tag{34}$$

- 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,

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla^\top f(\mathbf{x})(\mathbf{y}-\mathbf{x}) + \frac{L}{2}\|\mathbf{y}-\mathbf{x}\|^2 \tag{34}$$

- While we showed (34) assuming f is doubly differentiable, (34) 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 condition for Lipschitz continuity:

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \nabla^T f(\mathbf{x}^k)(\mathbf{x}^{k+1} - \mathbf{x}) + \frac{L}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 \quad \text{by (34)}$$

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,

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) + \frac{L}{2}\left\|\mathbf{x}^{k+1} - \mathbf{x}^k\right\|^2 \tag{35}$$

Convergence of Descent Algorithms: Generic and Specific Cases

Back to: Generic Convergence of Descent Algorithm Under strong wolfe..

- 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 $\{\mathbf{x}|f(\mathbf{x}) \leq f(\mathbf{x}^0)\}$
  - ∇f is Lipschiz continuous.

  Then, $\sum_{k=1}^{\infty} \frac{(\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k)^2}{\|\Delta\mathbf{x}^k\|^2} < \infty$ (that is, it is finite) < lim k-> inf [f(x^k) - f(x^0)]

  Proof:
- For any descent algorithm: $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$ for each k with each step: $\mathbf{x}^{k+1} = \mathbf{x}^k + t^k\Delta\mathbf{x}^k$.
- From the second Strong Wolfe condition:

$$\left|\nabla^T f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k)\Delta\mathbf{x}^k\right| \leq c_2\left|\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k\right| \tag{36}$$

Proving Convergence of Descent Algorithm

- Since $c_2 > 0$ and $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k < 0$,

$$\nabla^T f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k)\Delta\mathbf{x}^k \geq c_2\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{37}$$

- Subtracting $\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k$ from both sides of (37)

$$\left[\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\right]^T \Delta\mathbf{x}^k \geq (c_2 - 1)\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k \tag{38}$$

- By Cauchy Shwarz inequality and from Lipschitz continuity,

$$\left[\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\right]^T \Delta\mathbf{x}^k \leq \|\nabla f(\mathbf{x}^k + t^k\Delta\mathbf{x}^k) - \nabla f(\mathbf{x}^k)\|\|\Delta\mathbf{x}^k\| \leq L\|\Delta\mathbf{x}^k\|^2 t^k \tag{39}$$

Proving Convergence of Descent Algorithm (contd.)

- Combining (38) and (39),

$$t^k \geq \frac{c_2 - 1}{L} \frac{\nabla^T f(\mathbf{x}^k) \Delta \mathbf{x}^k}{\|\Delta \mathbf{x}^k\|^2} \tag{40}$$

- Substituting (40) into the first Wolfe condition $f(\mathbf{x}^k + t\Delta \mathbf{x}^k) < f(\mathbf{x}^k) + c_1 t \nabla^T f(\mathbf{x}^k) \Delta \mathbf{x}^k$

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k) - c_1 \frac{1 - c_2}{L} \frac{\left(\nabla^T f(\mathbf{x}^k) \Delta \mathbf{x}^k\right)^2}{\|\Delta \mathbf{x}^k\|^2} \tag{41}$$

- Substituting $c = c_1 \frac{1-c_2}{L}$ and applying (41) recursively,

$$f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^0) - c \sum_{i=0}^{k} \frac{\left(\nabla^T f(\mathbf{x}^i) \Delta \mathbf{x}^i\right)^2}{\|\Delta \mathbf{x}^i\|^2} \tag{42}$$

Exactly what I had wanted to prove... that the sum is upper bounded

Proving Convergence of Descent Algorithm (contd.)

- Taking limits of (42) as $k \to \infty$,

$$\lim_{k\to\infty} c\sum_{i=0}^{k} \frac{\left(\nabla^T f(\mathbf{x}^i)\Delta\mathbf{x}^i\right)^2}{\|\Delta\mathbf{x}^i\|^2} \leq \lim_{k\to\infty} f(\mathbf{x}^0) - f(\mathbf{x}^{k+1}) \leq \infty \tag{43}$$

  where the last inequality is because the descent algorithm proceeds only if $f(\mathbf{x}^{k+1}) \leq f(\mathbf{x}^k)$, and we have assumed that f is bounded below in $\Re^n$. This proves finiteness of the summation

- Thus, $\lim_{k\to 0} \frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|} = 0.$

Could converge to a saddle point
Hence convexity is imp for sufficiency of gradient going to 0

- If we additionally assume that the descent direction is not orthogonal to the gradient, *i.e.*, $-\frac{\nabla^T f(\mathbf{x}^k)\Delta\mathbf{x}^k}{\|\Delta\mathbf{x}^k\|\|\nabla f(\mathbf{x}^k)\|} \geq \Gamma$ for some $\Gamma > 0$, then, we can show[5] that $\lim_{k\to 0} \|\nabla f(\mathbf{x}^k)\| = 0$
- This shows convergence for a generic descent algorithm. What we are more interested in however, is the **rate of convergence** of a descent algorithm.

[5] Making use of the Cauchy Schwarz inequality