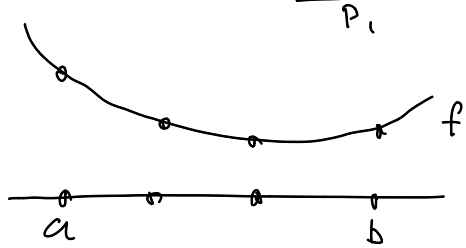
How to minimize a convex function over an interval in $\mathbb{R}$, $[a,b]$

Idea: evaluate f at a, $\underbrace{a+\frac{1}{3}(b-a)}_{p_1}$, $\underbrace{a+\frac{2}{3}(b-a)}_{p_2}$, b



If the minimum is at p_1, look in $[a, p_2]$

is at p_2, look in $[p_1, b]$

is at a, look in $[a, p_1]$

is at b, look in $[p_2, b]$

Decreases the width by $2/3$.

With more care (see Fibonacci or Golden Section Search) can reduce it by a factor of $\phi = \frac{1+\sqrt{5}}{2}$ per evaluation.

Minimizing smooth convex functions by gradient descent.

Recall from last lecture that for a convex function f:

1. x_* is a minimizer of f iff $\nabla f(x_*) = \bar{0}$
2. For all y and x, $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

The latter provides a useful lower bound on f.

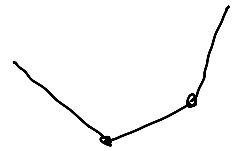
In this lecture, we will show that GD converges nicely if the gradients of f are "smooth."
That is, if they don't change too quickly.
We measure this in two ways.

1. We say that ∇f is L-Lipschitz if $\forall x, y$
$$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \|x-y\|_2$$

2. We consider the gradient of the gradient — the Hessian $\nabla^2 f$ which we recall is the matrix with entries $\left(\frac{\partial^2}{\partial x(i) \partial x(j)} f(x) \right)_{i,j}$

For small vectors δ, $\nabla f(x+\delta) - \nabla f(x) \approx \left(\nabla^2 f(x) \right) \delta$
So, we may upper bound changes in gradient by $\| \nabla^2 f(x) \|$

Note: neither approach can handle piecewise linear functions like



for which gradients are discontinuous.

They work well for $f(x) = \|Ax-b\|_2^2$,

which has $\nabla^2 f = A^T A$,

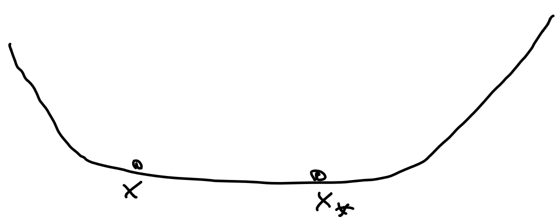
and $\nabla f(x) = 2A^T Ax - 2A^T b$,

so $\nabla f(x) - \nabla f(y) = 2A^T A(x-y)$

and $\|\nabla f(x) - \nabla f(y)\|_2 \leq 2\|A\|_2^2 \cdot \|x-y\|$,

So, it is $2\|A\|_2^2$ - Lipschitz

Note: we can have $f(x) \approx f(x_*)$, but x far from x_*:



Implications of L-Lipschitz: the gradient provides both upper and lower bounds.

The lower bound is $f(y) \geq f(x) + \nabla f(x)^T (y-x)$

<u>Lem 1</u> $\quad f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|_2^2$

<u>proof</u>

Let $z(t) = (1-t)x + ty$. $\quad z'(t) = y-x$

$\frac{d}{dt} f(z(t)) = z'(t) \cdot f'(z(t))$, so

$$f(y) - f(x) = f(z(1)) - f(z(0)) = \int_0^1 f'(z(t))\,dt$$

$$= \int_0^1 (y-x)^T \nabla f(z(t))\,dt$$

$$= \underbrace{\int_0^1 (y-x)^T \nabla f(x)\,dt}_{= (y-x)^T \nabla f(x)} + \underbrace{\int_0^1 (y-x)^T (\nabla f(z(t)) - \nabla f(x))\,dt}$$

$$\leq \int_0^1 \|y-x\| \, \|\nabla f(z(t)) - \nabla f(x)\|\,dt$$

$$\leq \int_0^1 \|y-x\| \cdot L \cdot \|t(y-x)\|\,dt$$

$$= \frac{1}{2} L \|y-x\|^2$$

<u>**Cor 1**</u> $\quad f(x) \leq f(x_*) + \frac{L}{2} \|x - x_*\|^2$

proof: apply lem 1 with $y = x$, $x = x_*$, $\nabla f(x_*) = \bar{0}$.

Gradient Descent: move from x to

$\hat{x} = x - \eta \nabla f(x)$, where η is the "step size".

Can choose η many ways. For $\eta = 1/L$ we show

<u>lemma 2</u> For $\hat{x} = x - \eta \nabla f(x)$, $\eta \le 1/L$

$$f(\hat{x}) \le f(x) - \frac{1}{2\eta} \|\hat{x} - x\|_2^2$$

<u>proof:</u>

Lem 1 =>

$$f(\hat{x}) \le f(x) + \nabla f(x)^T(-\eta \nabla f(x)) + \tfrac{1}{2} L \|\eta \nabla f(x)\|_2^2$$

$$= f(x) - \left(\eta - \frac{L\eta^2}{2}\right) \|\nabla f(x)\|_2^2$$

$$= f(x) - \eta\left(1 - \frac{L\eta}{2}\right) \|\nabla f(x)\|_2^2$$

$$\le f(x) - \tfrac{1}{2}\eta \|\nabla f(x)\|_2^2 \quad (\text{using } \eta \le 1/L)$$

$$= f(x) - \frac{1}{2\eta} \|\hat{x} - x\|_2^2$$

<u>lemma 3</u> For $\hat{x} = x - \eta \nabla f(x)$ with $\eta = 1/L$

$$f(\hat{x}) - f(x_*) \le \frac{L}{2}\left(\|x - x_*\|_2^2 - \|\hat{x} - x_*\|_2^2\right) \quad (2)$$

<u>proof</u> Lem 2 =>

$$f(\hat{x}) - f(x_*) \le f(x) - f(x_*) - \frac{L}{2} \|\hat{x} - x_*\|_2^2$$

Cor 1 =>

$$f(\hat{x}) - f(x_*) \le \frac{L}{2} \|x - x_*\| - \frac{L}{2} \|\hat{x} - x_*\|_2^2$$

Thm1

Let x_0 be the initial vector, and x_k be vector after k steps. Then for $\eta = 1/L$,

$$f(x_k) - f(x_*) \leq \frac{L}{2k} \|x_0 - x_*\|_2^2$$

Proof: Summing Lem 3 gives

$$\sum_{i=1}^{k} \left(f(x_i) - f(x_*) \right) \leq \frac{L}{2} \sum_{i=1}^{k} \|x_i - x_*\|_2^2 - \|x_{i-1} - x_*\|_2^2$$

$$= \frac{L}{2} \left(\|x_k - x_*\|_2^2 - \|x_0 - x_*\|_2^2 \right)$$

$$\leq \frac{L}{2} \|x_k - x_*\|_2^2$$

As $f(x_i)$ is motonically decreasing in i, $f(x_k) - f(x_*)$ is the smallest term, and so

$$f(x_k) - f(x_*) \leq \frac{L}{2k} \|x_k - x_*\|_2^2 .$$

So this converges, if not quickly.
Of course, choosing the optimal η at every step improves.

We can get faster convergence if we assume more.

f is m-strongly convex if $\sigma_n(\nabla^2 f(x)) \geq m$ for all x.

If g is any convex function, $g(x) + \frac{m}{2}\|x\|_2^2$ is m-strongly convex.

<u>Thm 2</u> for such f with $\eta = \frac{1}{L}$

$$f(x_k) - f(x_*) \leq \left(1 - \frac{m}{L}\right)^k \left(f(x_0) - f(x_*)\right)$$

This generalizes our analysis for least squares / lin equations.

<u>Lem 4</u> For all x, y

$$\frac{m}{2}\|y-x\|_2^2 \leq \left[f(y) - f(x) - \nabla f(x)^T (y-x)\right] \qquad (3)$$

<u>proof</u>

let $h(t) = f(tx + (1-t)y)$. Lagrange's form of Taylor's theorem gives $h(1) = h(0) + h'(0) + \frac{1}{2}h''(t)$ for some $t \in (0,1)$.

So $\exists z$ on $\overline{xy}$ s.t.

$$f(y) - f(x) - \nabla f(x)^T (y-x) = \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x),$$

and this latter term is at least $\frac{m}{2}\|y-x\|_2^2$

<u>Cor 2</u> As $\nabla f(x_*) = 0$, Lem 4 $\Rightarrow$ $f(x) - f(x_*) \geq \frac{m}{2}\|x - x_*\|_2^2$

So, x far from x_* $\Rightarrow$ $f(x)$ far from $f(x_*)$

We can also obtain an upper bound

Cor 3 $f(x) - f(x_*) \le \frac{1}{2m} \|\nabla f(x)\|_2^2$

Note: combining with Cor 2 gives $\|x - x_*\|_2^2 \le \frac{1}{4m^2} \|\nabla f(x)\|_2^2$

Proof:

$(3) \Rightarrow \forall x, y, \quad f(x) - f(y) \le \nabla f(x)^T (y-x) - \frac{m}{2}\|y-x\|_2^2$

Let $g = \nabla f(x)$, $z = y - x$, and consider $g^T z - \frac{m}{2}\|z\|^2$

$\nabla_z \left(g^T z - \frac{m}{2}\|z\|^2 \right) = g - mz$, so this is maximized when $z = \frac{1}{m} g$, at which point its value is $\frac{1}{2m}\|g\|^2$

So, $\forall x, y$ $f(x) - f(y) \le \frac{1}{2m} \|\nabla f(x)\|^2$.

We apply this with $y = x_*$

Proof of Theorem 2

Now, consider setting $\hat{x} = x - \eta \nabla f(x)$

We should pick η to minimize $f(\hat{x})$.

To prove such an η exists, we show $\eta = \frac{1}{L}$ is OK.

This choice gives

$$
\begin{aligned}
f(\hat{x}) &\le f(x) + \nabla f(x)^T (\hat{x} - x) + \frac{L}{2}\|\hat{x} - x\|_2^2 \\
&= f(x) - \frac{1}{L}\|\nabla f(x)\|_2^2 + \frac{1}{2L}\|\nabla f(x)\|_2^2 \\
&= f(x) - \frac{1}{2L}\|\nabla f(x)\|_2^2
\end{aligned}
$$

Cor 3 gives $f(\hat{x}) \le f(x) - \frac{m}{L}\left(f(x) - f(x_*)\right)$

$\Rightarrow f(\hat{x}) - f(x_*) \le \left(f(x) - f(x_*)\right)\left(1 - \frac{m}{L}\right)$