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

**Idea:** evaluate \( f \) at \( a, a + \frac{1}{3}(b-a), a + \frac{2}{3}(b-a), b \)

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 \( \frac{2}{3} \).

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 $\nabla f(x^*) = 0$
2. For all $y$ and $x$, $f(y) = f(x) + \nabla f(x)^T (y-x)$

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 $\nabla f$ 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
   $$\frac{\partial^2}{\partial x_i \partial x_j} f(x)$$

   For small vectors $\delta$, $\nabla f(x+\delta) - \nabla f(x) \approx (\nabla^2 f(y)) \delta$
   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 $\nabla^2 f = A^TA$,
and $\nabla f(x) = 2A^TAx - 2A^Tb$,
so $\nabla f(x) - \nabla f(y) = 2A^TA(x-y)$
and $\|\nabla f(x) - \nabla f(y)\|_2 \leq 2\|A\|_2^2 \cdot \|x-y\|_2$.
So, it is $2\|A\|_2^2$-Lipschitz.

Note: we can have $f(x) = f(x^*)$, but $x$ far from $x^*$.
Implications of L-Lipschitz: the gradient provides both upper and lower bounds.

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

**Lemma** \( f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{1}{2} \| y-x \|^2 \)

**Proof**

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

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

\[
\begin{align*}
f(y) - f(x) &= f(z(0)) - f(z(1)) = \int_0^1 \nabla f(z(t)) dt \\
&= \int_0^1 (y-x)^T \nabla f(z(t)) dt \\
&= \int_0^1 (y-x)^T \nabla f(z(t)) dt + \int_0^1 (y-x)^T (\nabla f(z(t)) - \nabla f(x)) dt \\
&\leq \int_0^1 \| y-x \| \| \nabla f(z(t)) - \nabla f(x) \| dt \\
&\leq \int_0^1 \| y-x \| \cdot L \cdot \| z(t)(y-x) \| dt \\
&= \frac{1}{2} L \| y-x \|^2
\end{align*}
\]

**Corollary** \( f(x) \leq f(x_*) + \frac{1}{2} \| x-x_* \|^2 \)

**Proof:** apply Lemma with \( y=x \), \( x=x_* \), \( \nabla f(x_*) = 0 \).
Gradient Descent: move from $x$ to
$$\hat{x} = x - \nabla f(x),$$ where $\nabla$ is the "step size".

Can choose $\nabla$ many ways. For $\nabla = \frac{1}{L}$ we show

**Lemma 2** For $\hat{x} = x - \nabla Df(x)$, $\nabla = \frac{1}{L}$
$$f(\hat{x}) \leq f(x) - \frac{1}{2 \nabla} \|\hat{x} - x\|_2^2$$

**proof:**

LEM 1 $\Rightarrow$
$$f(\hat{x}) \leq f(x) + Df(x)^T (x - \frac{1}{L} Df(x)) + \frac{1}{2} \|\frac{1}{L} Df(x)\|_2^2$$

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

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

$$\leq f(x) - \frac{1}{2} \nabla \|Df(x)\|_2^2 \quad \text{(using $\nabla \leq \frac{1}{L}$)}$$

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

**Lemma 3** For $\hat{x} = x - \nabla Df(x)$ with $\nabla = \frac{1}{L}$
$$f(\hat{x}) - f(x^*) \leq \frac{L}{2} \left(\|x - x^*\|_2^2 - \|\hat{x} - x^*\|_2^2\right) \quad (\star)$$

**proof** LEM 2 $\Rightarrow$
$$f(\hat{x}) - f(x^*) \leq f(x) - f(x^*) - \frac{L}{2} \|\hat{x} - x^*\|_2^2$$

LEM 1 $\Rightarrow$
$$f(\hat{x}) - f(x^*) \leq \frac{L}{2} \|x - x^*\|_2^2 - \frac{L}{2} \|\hat{x} - x^*\|_2^2$$
Thm 1

Let $x_0$ be the initial vector, and $x_k$ be vector after $k$ steps. Then for $L = \|
abla f(x_0)\|_2$, 

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

Proof: Summing lem 3 gives

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

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

$$= \frac{L}{2} \|x_k - x_\star\|_2^2.$$ 

As $f(x_i)$ is monotonically decreasing in $i$, 

$f(x_k) - f(x_\star)$ is the smallest term, and so

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

So this converges, if not quickly. 

Of course, choosing the optimal $n$ at every step improves.
We can get faster convergence if we assume more.

A function $f$ is $m$-strongly convex if $\sigma_0(\nabla^2 f(x)) \geq m$ for all $x$.

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

**Thm 2.** For such $f$ with $\sigma = \frac{1}{L}$,

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

This generalizes our analysis for least squares / linear equations.

**Lem 4.** For all $x, y$

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

**Proof.**

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

So, $\exists \zeta$ on $[t \theta] s.t.$

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

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

**Cor 2.** 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** \[ f(x) - f(x^*) \leq \frac{1}{2m} \|\nabla f(x)\|^2 \]

Note: Combining with Cor 2 gives \( \|x - x^*\|_2 \leq \frac{1}{4m^2} \|\nabla f(x)\|^2 \)

**proof:**

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

Let \( g = \nabla f(x) \), \( z = y - x \), and consider \( g^T z - \frac{m}{2} \|z\|_2^2 \)

\[ \nabla_v (g^T z - \frac{m}{2} \|z\|_2^2) = g - mz, \] so this is maximized when \( z = \frac{1}{m} g \), at which point its value is \( \frac{1}{2m} \|g\|_2^2 \)

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

We apply this with \( y = x^* \)

**Proof of Theorem 2**

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

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

To prove such an \( \eta \) exists, we show \( \eta = \frac{1}{\nabla f(x)} \) is OK.

This choice gives

\[ f(\hat{x}) \leq f(x) + \nabla f(x)^T(\hat{x} - x) + \frac{L}{2} \|\hat{x} - x\|_2^2 \]

\[ = f(x) - \frac{1}{2} \|\nabla f(x)\|_2^2 + \frac{1}{2L} \|\nabla f(x)\|^2 \]

\[ = f(x) - \frac{1}{2L} \|\nabla f(x)\|^2 \]

Cor 3 gives \( f(\hat{x}) \leq f(x) - \frac{m}{L} \left( f(x) - f(x^*) \right) \)

\[ \Rightarrow f(\hat{x}) - f(x^*) \leq (f(x) - f(x^*)) \left( 1 - \frac{m}{L} \right) \]