Today:
- The importance of convexity
- Gradient Descent + convergence rates
- How fn structure helps.

Main Qs:
- Why is convexity useful?
- How to exploit it algorithmically?

Let $f: \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. Then,

$$
\forall x, x_0 \in \mathbb{R}^d
$$

$$
f(x) \geq f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle
$$

(Linear fn of $x$ [e.g. $f(x) = 0.5 (a^T x)$])

Hence, the 1-st order Taylor expansion of $f$ is a "global underestimate"
Observe: 1st order Taylor always has a linear form, e.g., \( ax + b \)

Important Remark:

What happens for \( x_0 \) s.t. \( \nabla f(x_0) = 0 \)?

From 1st order Taylor:

\[
\begin{align*}
f(x) &\approx f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle \\
\Rightarrow f(x) &\approx f(x_0) & \forall x \in \mathbb{R}
\end{align*}
\]

That is, all points \( x_0 \in \mathbb{R} \) s.t. \( \nabla f(x_0) = 0 \) are global minimizers of \( f \).
Q: Can we use the linear approx property to devise an algorithm for
\[
\min_{x \in \mathbb{R}} f(x)
\]
whe \( f(.) \) is \( \nabla x \)?

Idea:

Say we "initialize" at \( x_0 \) we could try to follow \( f(x_0) + \nabla f(x_0)(x-x_0) \), but for how long?

If we tried to minimize it would lead us to a point that is not a min for \( f(x) \) since the linear approx is unbounded.

Clue: Take a small step!
Let's assume that our algorithmic progress is captured by:

\[ x_{k+1} = x_k + u_k \]

**Goal:** Reach \( \| \nabla f(x_k) \| \to 0 \)

Use "small step" clue from above:

\[ x_{k+1} = \arg \min_x \left\{ f(x_k) + \langle \nabla f(x_k), x-x_k \rangle + \frac{1}{2\gamma} \| x-x_k \|_2^2 \right\} \]

\( \gamma \) mini. the Lin. approximation but not too much!

Think of \( x_{k+1} \) as the "best \( x \) near \( x_k \)",

What is the optimal solution of the above "local" optimization?

Observe that it is a quadratic function in \( x \)!
Set the gradient of the above function to 0.

\[ \nabla_x \left\{ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \| x - x_k \|_2^2 \right\} = 0 \]

\[ \Rightarrow \nabla f(x_k) + \frac{1}{\gamma} (x - x_k) = 0 \]

\[ \Rightarrow x_{k+1} = x_k - \gamma \nabla f(x_k) \]

Gradient Descent is minimizing

\[ f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \| x - x_k \|_2^2 \]

at every step.

How fast does G{D} converge?

Short answer: It depends on the function!
Convergence rates:

General form:
\[ f(x_t) - f^* \leq p(T) \cdot s(f) \]

Distance to optimum

Convergence rates tell us how fast we approach a (locally) optimal solution, in the worst case w.r.t. all functions in the class we are analyzing.

\[ f(x_t) - f^* \]
\[ p(T) \cdot s(f) \] is an upper bound on performance

\[ T = \# \text{ iterations} \]

Warning: true performance may be significantly better.
Warning 2: if Alg1 has faster conv rates than Alg2 that doesn’t imply Alg1 is actually faster.

However, rates are informative and can help us understand what structures allow for faster algorithms, and sometimes can be good guides towards algorithm design.

Let’s see an example of conv rates:

Lipschitz functions:

\[ |f(x) - f(y)| \leq \|x - y\| \]

\[ \|\nabla f(x)\| \leq L \quad \forall x \in X \]
**Theorem:** Let $f$ be convex and assume that $\|x_1 - x_2\| \leq R$ and $\|\nabla f(x)\| \leq L$ (which implies $L$-Lip.).

Then, if we set $\gamma = \frac{R}{L \sqrt{T}}$

$$f\left(\frac{1}{T} \sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{RL}{\sqrt{T}}$$

**Proof:**

We start with the under estimator:

$$f(x_k) - f(x^*) \leq \langle \nabla f(x_k), x_k - x^* \rangle$$

$$= \frac{\langle x_k - x_{k+1}, x_k - x^* \rangle}{\gamma}$$

We will use the following

$$a^T b = \|a\|^2 + \|b\|^2 - \|a - b\|^2$$

$$\Rightarrow f(x_k) - f^* \leq \frac{1}{2\gamma} \left\{ \|x_k - x^*\|^2 + \|x_k - x_{k+1}\|^2 - \|x_{k+1} - x^*\|^2 \right\}$$
Hence,

\[
f(x_k) - f^* \leq \frac{\gamma}{2\delta} \sum_{t=1}^{T} \|x_k - x^*\|^2 - \|x_k - x^*\|^2 \sum_{t=1}^{T} \frac{1}{\gamma} \|f'(x_t)\|^2 \leq L^2
\]

\[
f(x_k) - f^* \leq \frac{\gamma}{2\delta} \|x_{k-1} - x^*\|^2 - \|x_k - x^*\|^2 \sum_{t=1}^{T} \frac{1}{\gamma} \|f'(x_t)\|^2 \leq L^2
\]

\[\vdots\]

\[
f(x_0) - f^* \leq \frac{\gamma}{2\delta} \|x_0 - x^*\|^2 - \|x_1 - x^*\|^2 \leq \frac{\gamma}{2} L^2
\]

Sum all the above:

\[
\sum_{t=1}^{T} (f(x_t) - f^*) \leq -\|x_{T+1} - x^*\|^2 + \|x_0 - x^*\|^2 \frac{2\delta}{\gamma} + T \frac{\gamma L^2}{2}
\]

\[
\Rightarrow \frac{1}{T} \sum_{t=1}^{T} f(x_t) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\delta T} + \frac{\gamma L^2}{2}
\]

\[
\Rightarrow \text{Due to convexity:}
\]

\[
f\left(\frac{1}{T} \sum_{t=1}^{T} x_t\right) - f^* \leq \frac{1}{T} \sum_{t=1}^{T} f(x_t) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\delta T} + \frac{\gamma L^2}{2}
\]
Find the best $\gamma$:

$$\min_{\gamma} \frac{R^2}{2\gamma^2} + \frac{xL^2}{2} \implies \gamma = \frac{R}{L\sqrt{T}}$$

$$\implies f\left(\frac{4L}{T}x_t\right) - f^* \leq \frac{R\cdot L}{\sqrt{T}}$$

Cor. For $\varepsilon$-approx we need

$$\varepsilon = \frac{R\cdot L}{\sqrt{T}} \implies T = \frac{R^2L^2}{\varepsilon^2}$$ steps.

Main message: structure helps!

Next time: Performance of GD on
- Smooth
- Str. CVx
- nonconvex functions
- Complexity to reach $\varepsilon$-accuracy.