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

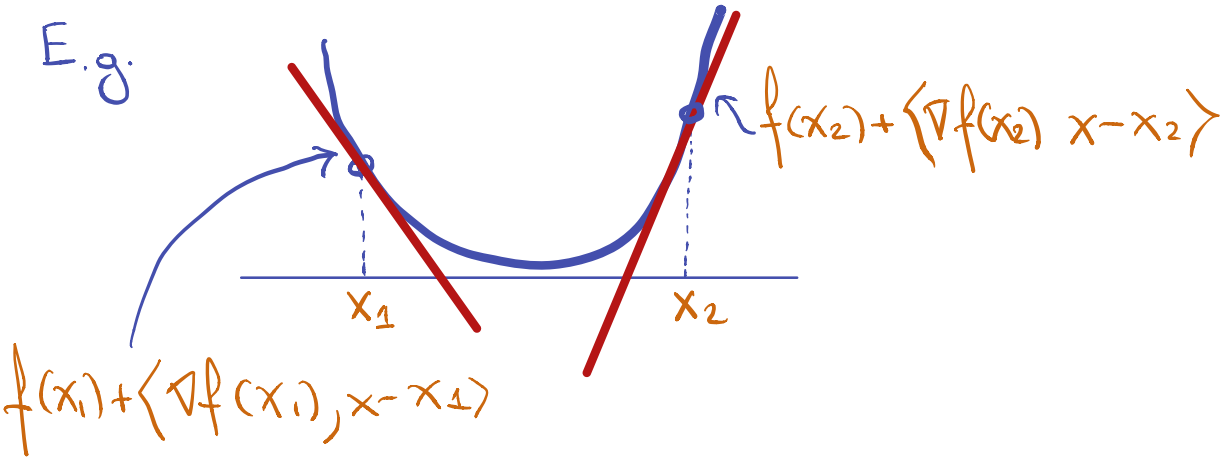
$$\forall x, x_0 \in \mathcal{X}^d$$

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

linear fn of x [e.g $f(x) = \Theta(a^T x)$]

Hence, the 1-st order Taylor expansion of f is a "global underestimator"

E.g.



Observe: 1-st order taylor always has a linear form, e.g., $a^T x + b$

Important Remark:

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

From 1st order Taylor:

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

$$\Rightarrow f(x) \geq f(x_0) \quad \forall\ x \in \mathcal{X}$$

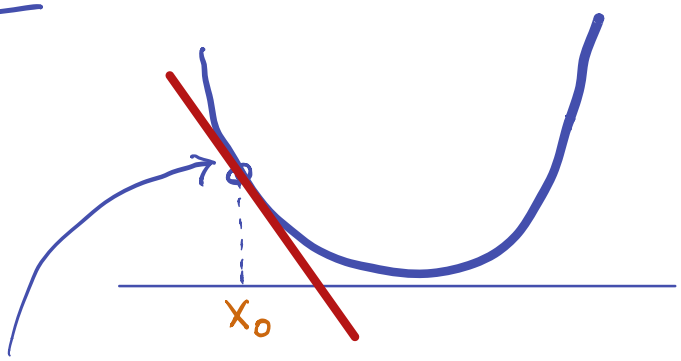
That is, all points $x_0 \in \mathcal{X}$ s.t. $\nabla f(x_0)=0$ are global minimizers of f

Q: Can we use the linear approx. property to devise an algorithm for

$$\min_{x \in \mathcal{X}} f(x)$$

when $f(\cdot)$ is cvx ?

Idea:



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

If we tried to minimize $f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle$ 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:

$$x_{k+1} = x_k + d_k$$

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

Use "small step" clue from above:

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

↳ min. 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 fn in x!

Set the gradient of the above fn to 0.

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

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

$$\Rightarrow \boxed{x_{k+1} = x_k - \gamma \cdot \nabla f(x_k)}$$

Gradient Descent is minimizing

$$f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\gamma} \|x - x_k\|^2$$

at every step.

How fast does GD converge?

Short answer: It depends on the function!

Convergence rates:

General form:

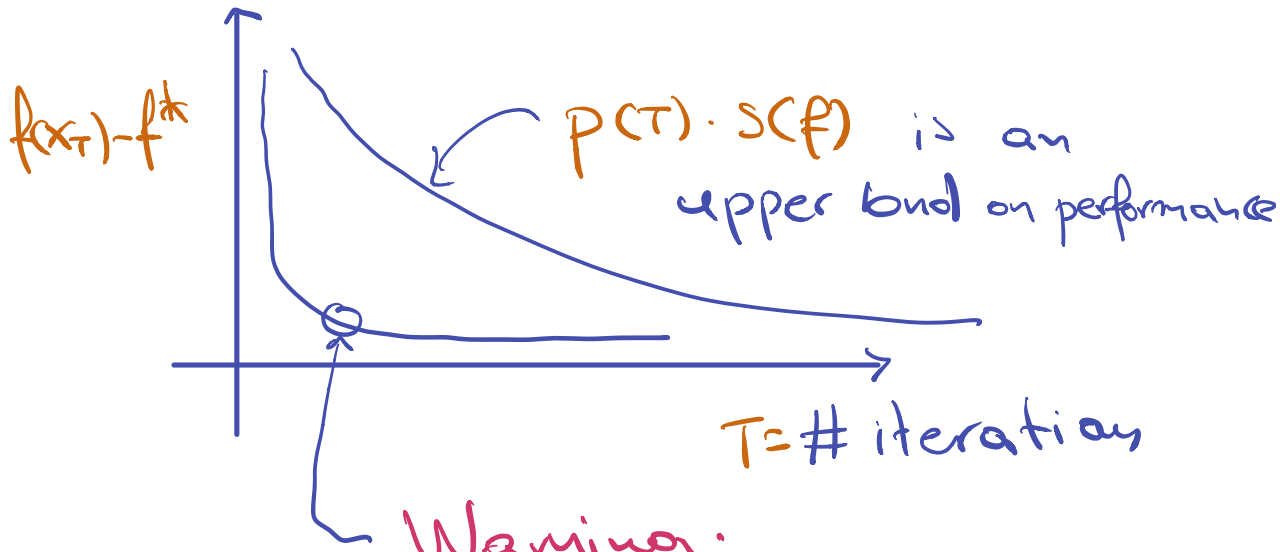
$$f(x_T) - f^* \leq p(T) \cdot s(f)$$

$f(x_T) - f^*$: Distance to optimum

$p(T)$: "rate" function

$s(f)$: depends on $f(\cdot)$

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.



Warning: true performance may be significantly better.

Warning 2: if Alg 1 has faster conv rates than Alg 2 that doesn't imply Alg 1 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:

$$|f(x) - f(y)| < L\|x - y\|$$

$$\Leftarrow \quad \|\nabla f(x)\| \leq L \quad \forall\, x \in \mathcal{X}$$

Theorem: Let f be convex and assume that $\|x_1 - x^*\| \le R$ and $\|\nabla f(x)\| \le L$ (which implies L-Lip.)

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

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

Proof:

We start with the under estimator:

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

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

We will use the following

$$a^T b = \frac{\|a\|^2 + \|b\|^2 - \|a-b\|^2}{2}$$

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

Hence,

$$f(x_k)-f^* \le \frac{1}{2\gamma}\left\{\|x_k-x^*\|^2-\|x_{k+1}-x^*\|^2\right\}+\frac{\gamma}{2}\underbrace{\|\nabla f(x_k)\|^2}_{\le L^2}$$

$$f(x_{k-1})-f^* \le \frac{1}{2\gamma}\left\{\|x_{k-1}-x^*\|^2-\|x_k-x^*\|^2\right\}+\frac{\gamma}{2}L^2$$

$$\vdots$$

$$f(x_0)-f^* \le \frac{1}{2\gamma}\left\{\|x_0-x^*\|^2-\|x_1-x^*\|^2\right\}+\frac{\gamma}{2}L^2$$

Sum all the above:

$$\sum_{t=1}^{T}\left(f(x_t)-f^*\right) \le \frac{-\|x_{T+1}-x^*\|^2+\|x_0-x^*\|^2}{2\gamma} + T\cdot\frac{\gamma L^2}{2}$$

$$\Rightarrow \frac{1}{T}\sum_t f(x_t)-f^* \le \frac{\|x-x^*\|^2}{2\gamma T}+\frac{\gamma L^2}{2}$$

$\Rightarrow$ Due to convexity:

$$f\left(\frac{1}{T}\sum_t x_t\right)-f^* \le \frac{1}{T}\sum_t f(x_t)-f^* \le \frac{R^2}{2\gamma T}+\frac{\gamma L^2}{2}$$

Find the best γ:

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

$$\Rightarrow f\left(\frac{1}{T}\sum_t x_t\right) - f^* \le \frac{R \cdot L}{\sqrt{T}}$$

$\blacksquare$

Cor. For ε-approx we need

$$\varepsilon = \frac{RL}{\sqrt{T}} \Rightarrow T = \frac{R^2 L^2}{\varepsilon^2} \text{ steps.}$$

Main message: structure helps!

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