No class tuesday 11/11
Exam friday 11/14

CS7800: Advanced Algorithms

Class 18: Convex Optimization

- Basic Concepts
- Gradient Descent

Jonathan Ullman
November 7, 2025

Exam 2:

- Linear programming
- Reductions
  - NP-completeness/hardness
  - Applications of max flow / min cut

Convex Optimization

Linear Programming

$$\min \; c^T x$$
$$\text{s.t.} \; Ax \geq b$$

min/max interchangeable

linear function

linear constraints

⟹

Convex Programming

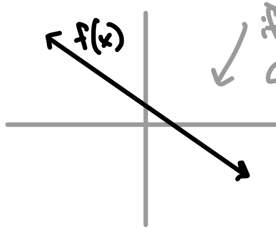
$$\min \; f(x)$$
$$\text{s.t.} \; x \in S$$

min important

convex function

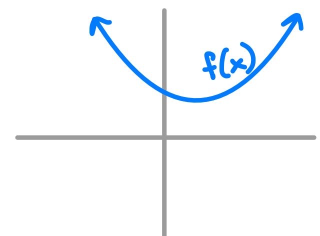
convex set

Objective



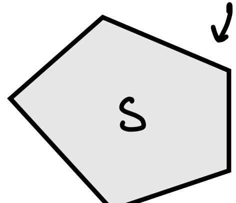
only interesting if there are constraints

⟹



interesting even without constraints

Feasible region



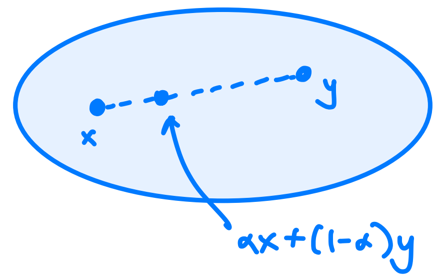
polytope

⟹

convex set

Convex Sets

A <u>set</u> $S \subseteq \mathbb{R}^n$ is <u>convex</u> if for every $x, y \in S$ and $0 \le \alpha \le 1$ we have $\alpha x + (1-\alpha) y \in S$

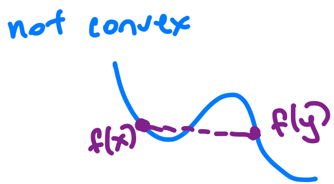


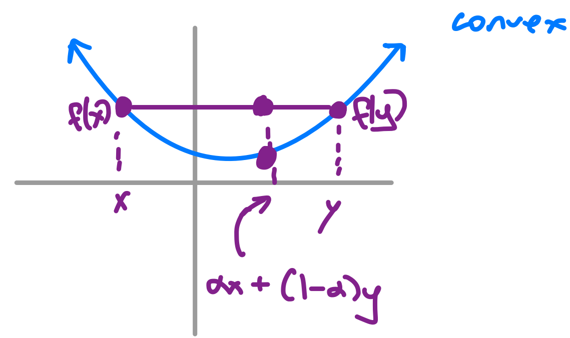
<u>Examples:</u>

- $S = \{x : \|x\| \le B\}$ for any norm B ← feasibility easy
- $S = \{x : Ax \le b$ and $x \ge 0$ for any $A, b\}$ ← feasibility tricky

Convex Functions

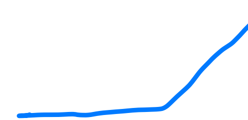
A <u>function</u> $f: \mathbb{R}^n \to \mathbb{R}$ is <u>convex</u> if for every $x, y \in \mathbb{R}^n$ and $0 \le \alpha \le 1$ we have $f(\alpha x + (1-\alpha) y) \le \alpha f(x) + (1-\alpha) f(y)$





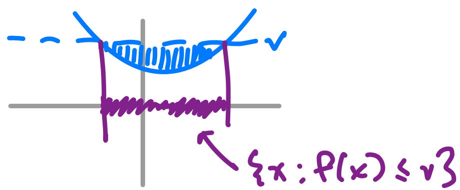
<u>Examples:</u>

- $f(x) = \|x\|$ for any norm
- $\|Ax - b\|_2^2$ ⟵ ordinary least squares regression
- Linear functions
- $\max\{0, x\}$ ⟵ called ReLu or Hinge loss

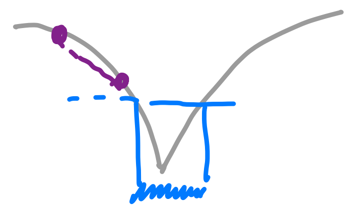


Properties of Convex Functions

If f is a convex function then $\{x : f(x) \leq v\}$ is a convex set



Note the reverse is not true!



Properties of Convex Functions

If f is a differentiable convex function then for any $x, y \in \mathbb{R}^n$

$$f(y) \geq \underbrace{f(x) + \langle \underbrace{\nabla f(x)}_{\text{gradient of } f \text{ at } x}, y - x \rangle}_{}$$

tangent line to f at x

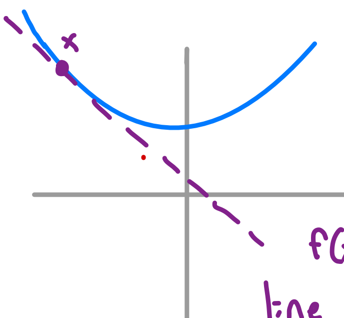
gradient of f at x

Gradient operator

For a differentiable $f : \mathbb{R}^n \to \mathbb{R}$

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_n}(x) \right)$$

One dimensional intuition



$f(x) = ax^2 + bx + c$

$f'(x) = 2ax + b$

$f(x) + f'(x) \cdot (y - x) = \ell(y)$

line tangent to f at x

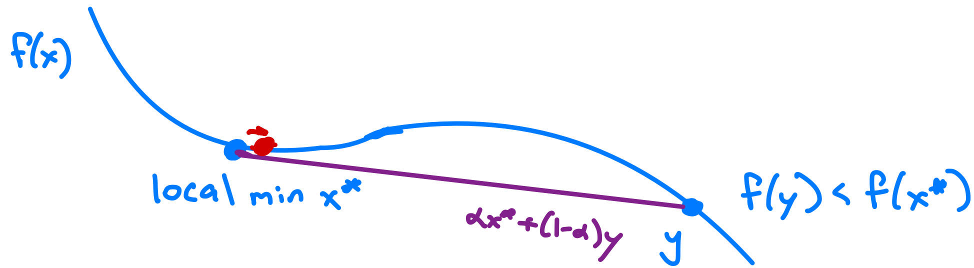
$f(x) = x^T A x + b^T x + c$

$\nabla f(x) = 2Ax + b^T$

Properties of Convex Functions

Thm: Any local minimum of a convex function is a global minimum

Proof:

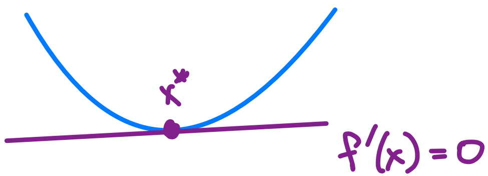


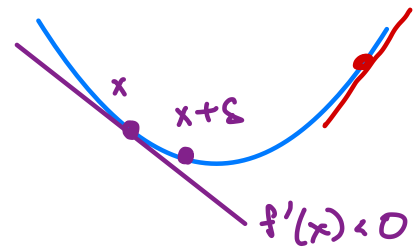
Properties of Convex Functions

Thm: If f is differentiable and convex then x^* is a global minimum if and only if $\langle \nabla f(x^*), x - x^* \rangle \geq 0$ for every $x \in S$

If $S = \mathbb{R}^n$ then this is $\nabla f(x) = 0$

One dimensional picture

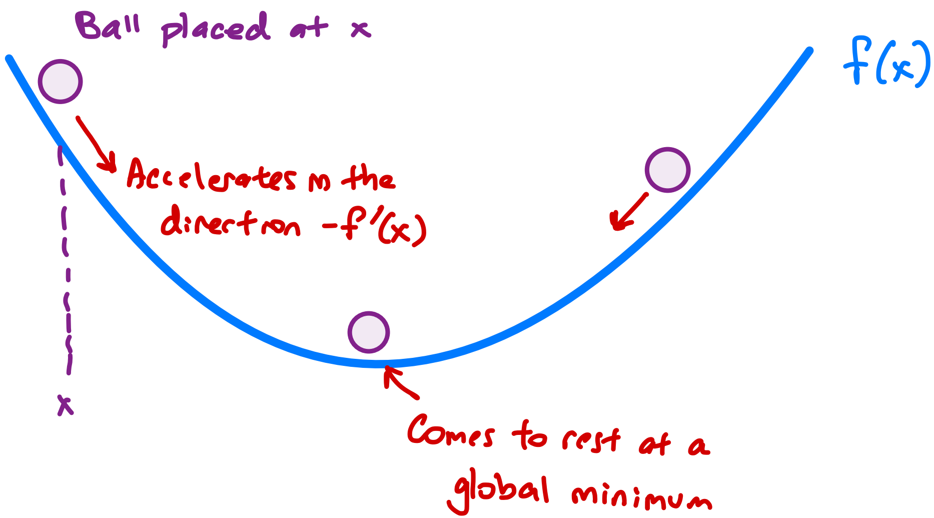




Theorem $\Rightarrow$ If x is not a global minimum then there exists a point y such that $\langle \nabla f(x), y - x \rangle < 0$, so moving in the direction $y - x$ will decrease the function

If $S = \mathbb{R}^n$ then $y = -\nabla f(x)$

Convex Optimization Intuition



In $\mathbb{R}^n$, ball would accelerate in the direction $-\nabla f(x)$ and come to rest where $\nabla f(x) = 0$

First Order Convex Optimization

Linear Programming

$$\min \; c^T x$$
$$Ax \geq b$$

Input is: $A \in \mathbb{R}^{m \times n}$

$b \in \mathbb{R}^m$

$c \in \mathbb{R}^n$

Convex Programming

$$\min \; f(x)$$
$$x \in S$$

No general compact way to describe $f(x)$

Input is: An oracle that takes $x \in \mathbb{R}^n$ and returns $\nabla f(x)$

Gradient Descent

Simplest setting

- $f: \mathbb{R}^n \to \mathbb{R}$ differentiable
- $S = \mathbb{R}^n$ (unconstrained)

Choosing step size: need to balance making progress against "overshooting" the minimizer

① $\arg\min_{\eta} f(x^{(t-1)} - \eta \nabla f(x^{(t-1)}))$

② $\eta_t \approx 1/t$

③ $\underline{\eta_t = 1/\sqrt{T}}$

Gradient descent

Initialize $x^{(0)}$

For $t = 1, \dots, T$:

  choose step size η_t

  let $x^{(t)} = x^{\cancel{(t)}(t-1)} - \eta_t \nabla f(x^{(t-1)})$

Return $\frac{1}{T}\sum_{t=1}^{T} x^{(t)}$

In practice return $x^{(T)}$

Analyzing Gradient Descent

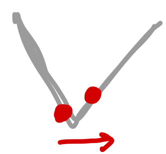
A potential function

How do we keep track of progress? Two natural choices:

① (Function value) $\Phi_t = f(x^{(t)}) - f(x^*)$

② (Distance to optimality) $\Phi_t = \|x^{(t)} - x^*\|_2^2$

Also combinations thereof



Analysis outline:

- Bound the decrease in potential $\Phi_{t-1} - \Phi_t \geq B_t$
- Telescoping sum gives $\Phi_T \leq \Phi_0 - \sum_{t=1}^{T} B_t$

Analyzing Gradient Descent

Function f

Iterates $x^{(t)} = x^{(t-1)} - \eta \nabla f(x^{(t-1)})$

Potential $\Phi_t = \frac{1}{2\eta}\|x^{(t)} - x^*\|_2^2$

$\|v\|^2 = \langle v, v\rangle$

$$\Phi_t - \Phi_{t-1}$$

$$= \frac{1}{2\eta}\left[\|x^{(t)} - x^*\|^2 - \|x^{(t-1)} - x^*\|^2\right]$$

$$= \frac{1}{2\eta}\left[\langle x^{(t)} - x^*, x^{(t)} - x^*\rangle - \langle x^{(t-1)} - x^*, x^{(t-1)} - x^*\rangle\right]$$

$$= \frac{1}{2\eta}\left[\langle x^{(t)} - x^{(t-1)}, x^{(t)} + x^{(t-1)} - 2x^*\rangle\right]$$

$x^{(t)} - x^{(t-1)} = -\eta \nabla f(x^{(t-1)})$

$$= \frac{1}{2\eta}\left[\langle -\eta \nabla f(x^{(t-1)}), -\eta \nabla f(x^{(t-1)}) + 2x^{(t-1)} - 2x^*\rangle\right]$$

$$= \frac{1}{2\eta}\left[\eta^2\|\nabla f(x^{(t-1)})\|^2 + 2\eta\langle \nabla f(x^{(t-1)}), x^* - x^{(t-1)}\rangle\right]$$

Analyzing Gradient Descent

$$\Phi_t - \Phi_{t-1}$$

$$= \frac{1}{2\eta}\left[\eta^2 \|\nabla f(x^{(t-1)})\|^2 + 2\eta \langle \nabla f(x^{(t-1)}), x^* - x^{(t-1)} \rangle\right]$$

$$= \frac{\eta}{2}\|\nabla f(x^{(t-1)})\|^2 + \langle \nabla f(x^{(t-1)}), x^* - x^{(t-1)} \rangle$$

$$f(x^*) - f(x^{(t-1)}) \geq \langle \nabla f(x^{(t-1)}), x^* - x^{(t-1)} \rangle$$

$$\leq \underbrace{\frac{\eta}{2}\|\nabla f(x^{(t-1)})\|^2}_{\text{Assume } \|\nabla f\|^2 \leq G^2} + \underbrace{f(x^*) - f(x^{(t-1)})}_{\leq 0}$$

$$\Phi_T - \Phi_0 = (\Phi_T - \Phi_{T-1}) + (\Phi_{T-1} - \Phi_{T-2}) + \ldots$$

$$\leq \frac{T\eta G^2}{2} + \sum_{t=1}^{T} f(x^*) - f(x^{(t-1)})$$

Analyzing Gradient Descent

$$\bar{x} = \frac{1}{T}\sum_{t=1}^{T} x^{(t)}$$

$$\Phi_T - \Phi_0 = (\Phi_T - \Phi_{T-1}) + (\Phi_{T-1} - \Phi_{T-2}) + \dots$$

$$\leq \frac{T\eta G^2}{2} + \sum_{t=1}^{T} f(x^*) - f(x^{(t-1)})$$

convexity

$$f(\bar{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x^{(t-1)}) - f(x^*) \leq \frac{\eta G^2}{2} + \frac{\Phi_0}{T} - \frac{\Phi_T}{T} \leq \frac{\eta G^2}{2} + \frac{\Phi_0}{T}$$

$$f(\bar{x}) - f(x^*) \leq \frac{\eta G^2}{2} + \frac{\|x^{(0)} - x^*\|^2}{2\eta T}$$

Assume $R^2 \geq \|x^{(0)} - x^*\|^2$

$$\leq \frac{\eta G^2}{2} + \frac{R^2}{2\eta T}$$

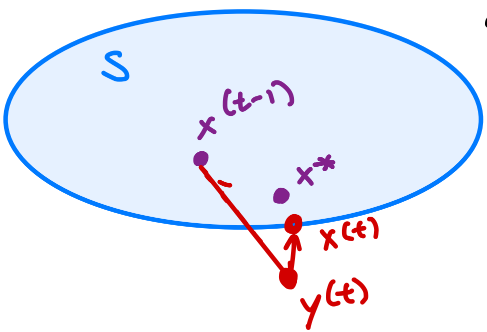
set η optimally

$$f(\bar{x}) - f(x^*) \leq \frac{RG}{\sqrt{T}}$$

Constrained Optimization

Simplest setting

- $f: \mathbb{R}^n \to \mathbb{R}$ differentiable
- $S \subseteq \mathbb{R}^n$



Ex: $S = \{x : \|x\|_2 = 1\}$

$$x^{(t)} = \frac{y^{(t)}}{\|y^{(t)}\|}$$

Thm: $\|x^{(t)} - x^*\|_2 \le \|y^{(t)} - x^*\|_2$

"Projection decreases distance"

Projected Gradient Descent

Initialize $x^{(0)}$

For $t = 1, \dots, T$:

choose step size η_t

let $y^{(t)} = x^{(t-1)} - \eta_t \nabla f(x^{(t-1)})$

$$\boxed{\text{let } x^{(t)} = \underset{x \in S}{\operatorname{argmin}} \|x - y^{(t)}\|_2}$$

Return $\frac{1}{T} \sum_{t=1}^{T} x^{(t)}$

"Projection" can often be computed efficiently

Stochastic Optimization

Often in machine learning and statistics

model parameters

$$f(x) = \frac{1}{n} \sum_{i=1}^{n} \ell_{z_i}(x)$$

error of model on example z_i

Computing $\nabla f(x)$ is expensive, so we use a single $\underbrace{\nabla \ell_z(x) \text{ for one random } z}_{\text{a stochastic gradient}}$

Fact:

$$\mathbb{E}_z[\nabla \ell_z(x)] = \nabla f(x)$$