CSE 5401 Lec. 9

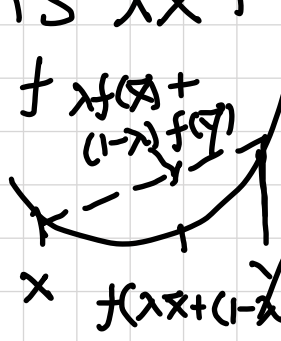
Gradient descent

1. Convex optimization
2. Gradient descent
3. Faster convergence for smooth & strongly convex functions

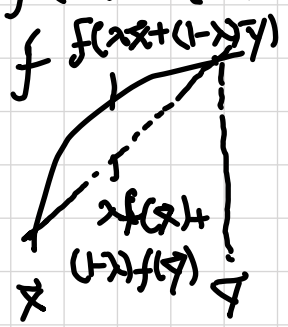
Algorithms for solving continuous optimization problems are versatile yet effective tools in the design of algorithms for other problems. Among these, convex optimization methods are generally the most broadly applicable methods with polynomially-bounded running time. Recall, a function $f: \mathbb{R}^n \to \mathbb{R}$ is <u>convex</u> if for every $\vec{x}, \vec{y} \in \mathbb{R}^n$ and $\lambda \in [0,1]$ the convex combinations $\lambda\vec{x} + (1-\lambda)\vec{y}$ satisfy $f(\lambda\vec{x}+(1-\lambda)\vec{y}) \le \lambda f(x) + (1-\lambda)f(y)$

In pictures:

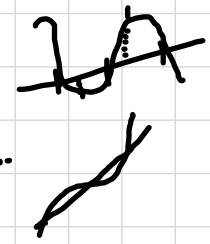
convex
x^2, e^x, etc.



<u>not</u> convex
(concave)
$-\sqrt{x}$, $\log x$, ...



neither
$\sin x$, $\cos x$, x^3, ...



Similarly, a set $K \subseteq \mathbb{R}^n$ is <u>convex</u> if for any $\vec{x}, \vec{y} \in K$, the convex combinations $\lambda\vec{x} + (1-\lambda)\vec{y} \in K$.

A convex optimization problem has the form, given the convex objective f and convex set K, find $\vec{x} \in K$ minimizing $f(\vec{x})$.

For example, linear programming is a convex optimization problem: if $\langle \vec{a}, \vec{x} \rangle \le b$ and $\langle \vec{a}, \vec{y} \rangle \le b$ $\langle \vec{a}, \lambda\vec{x} + (1-\lambda)\vec{y} \rangle =$
$\lambda \langle \vec{a}, \vec{x} \rangle + (1-\lambda) \langle \vec{a}, \vec{y} \rangle \le b$.

Furthermore, if $K_1, \dots, K_m$ are convex, so is $\bigcap_{i=1}^{m} K_i$.

But also, for any norm $\|\cdot\|$, $\|\lambda\vec{x} + (1-\lambda)\vec{y}\| \le \lambda\|\vec{x}\| + (1-\lambda)\|\vec{y}\|$

For any positive semidefinite matrix A ($\vec{x}^T A \vec{x} \ge 0 \ \forall \vec{x}$)

$f(\vec{x}) = \vec{x}^T A \vec{x}$ is convex

$f(\vec{x}) = \log(e^{x_1} + e^{x_2} + \cdots + e^{x_n})$ is convex

$\max\{f(\vec{x}), g(\vec{x})\}$ for convex f & g is convex etc.

Convex optimization is particularly nice for two main reasons

① Only global optima can be local optima — so local optimization techniques are effective

② Convex sets and functions can be usefully approximated by linear constraints and functions

We'll elaborate on both.

Proof of ①: If $\vec{x}$ is a local minimum of f on K, then for every $\vec{y}$ there is some $\varepsilon > 0$ such that

$f(\vec{x}) \le f((1-\varepsilon)\vec{x} + \varepsilon\vec{y}) \le (1-\varepsilon)f(\vec{x}) + \varepsilon f(\vec{y}) \Rightarrow \cancel{\varepsilon} f(\vec{x}) \le \cancel{\varepsilon} f(\vec{y})$

so $\vec{x}$ is a global optimum of f. ▨

Towards ②, convex functions always have subgradients

Defn: For $K \subseteq \mathbb{R}^n$, $\vec{g} \in \mathbb{R}^n$ is a subgradient of $f: K \to \mathbb{R}$ at $\vec{x}$ if for all $\vec{y} \in K$, $f(\vec{x}) - f(\vec{y}) \le \vec{g}^T(\vec{x} - \vec{y})$.

If f is differentiable at $\vec{x}$, the gradient ∇f is a subgradient

Proof: Since $f((1-\lambda)\vec{x} + \lambda\vec{y}) \le (1-\lambda)f(\vec{x}) + \lambda f(\vec{y})$, $\frac{f((1-\lambda)\vec{x} + \lambda\vec{y}) - (1-\lambda)f(\vec{x})}{\lambda} \le f(\vec{y})$

i.e., $f(\vec{x}) + \frac{f(\vec{x} + \lambda(\vec{y} - \vec{x})) - f(\vec{x})}{\lambda} \xrightarrow{\lambda \to 0} f(\vec{x}) + \nabla f(\vec{x})^T(\vec{y} - \vec{x}) \le f(\vec{y})$ ▨

It is also immediate that $\vec{x}$ is a global minimum if $\vec{0}$ is a subgradient of f.

In gradient descent, we seek to minimize a convex function f by iteratively following the gradient: for some small step size γ, $\vec{x}^{(i)} \leftarrow \vec{x}^{(i-1)} - \gamma \nabla f(\vec{x}^{(i-1)})$. Locally, $-\nabla f(\vec{x})$ is the direction of greatest decrease of f at $\vec{x}$. For constrained optimization, the gradient step may take us outside the set K. We then need to be able to find the projection of the point in order to remain in K:

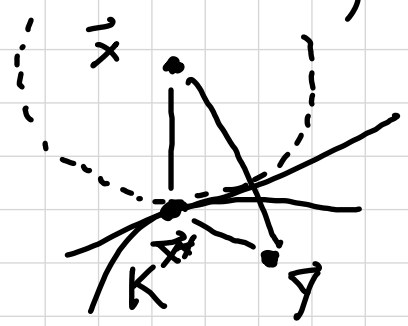
Defn The projection of $\vec{x}$ on K is $\vec{y} = \underset{\vec{y} \in K}{\operatorname{argmin}} \|\vec{x} - \vec{y}\|_2$.

Thus the algorithm is

For $t = 1, \ldots, T$
  $\vec{x}^{(t)} \leftarrow \text{project}(\vec{x}^{(t-1)} - \gamma \nabla f(\vec{x}^{(t-1)}), K)$
Return $\frac{1}{T} \sum_{t=0}^{T-1} \vec{x}^{(t)}$

Note that the gradient need not vanish if the optimum is attained at the boundary of K. For closed convex K and differentiable convex f, $\vec{x}^* = \underset{\vec{y} \in K}{\operatorname{argmin}} f(\vec{y}) \iff \forall \vec{y} \in K \; \nabla f(\vec{x}^*)^T (\vec{x}^* - \vec{y}) \leq 0$.
(It's immediate that if $\nabla f(\vec{x})^T (\vec{x} - \vec{y}) > 0$ for some $\vec{y} \in K$, then f decreases in the $\vec{y} - \vec{x}$ direction in K, so such $\vec{x}$ are not optimal; conversely, if $\nabla f(\vec{x})^T (\vec{x} - \vec{y}) \leq 0$ for all $\vec{y} \in K$, f is nondecreasing in all directions in K.)

In particular, from the definition of projection, $\nabla \|\vec{x}^* - \vec{x}\|_2^2 = 2(\vec{x}^* - \vec{x})$ so letting $\pi_K(\vec{x})$ denote the projection of $\vec{x}$ on K, $\forall \vec{y} \; (\pi_K(\vec{x}) - \vec{x})^T (\pi_K(\vec{x}) - \vec{y}) \leq 0$



Therefore also $\|\vec{y} - \vec{x}\|_2^2 = \|(\vec{y} - \pi_K(\vec{x})) + (\pi_K(\vec{x}) - \vec{x})\|_2^2 =$
$\|\vec{y} - \pi_K(\vec{x})\|_2^2 + 2\langle \vec{y} - \pi_K(\vec{x}), \pi_K(\vec{x}) - \vec{x} \rangle + \|\pi_K(\vec{x}) - \vec{x}\|_2^2$
$\geq \|\vec{y} - \pi_K(\vec{x})\|_2^2 + \|\pi_K(\vec{x}) - \vec{x}\|_2^2$. In particular, $\|\vec{y} - \pi_K(\vec{x})\|_2 \leq \|\vec{y} - \vec{x}\|_2$

We're now ready to analyze the algorithm.

Theorem Let K be a closed convex set and f be a convex function. Suppose all $\vec{y} \in K$ have (sub)gradient $\|\nabla f(\vec{y})\|_2 \le G$, and all $\vec{x}, \vec{y} \in K$ have $\|\vec{x} - \vec{y}\|_2 \le D$. Then projected gradient descent with $T = \frac{D^2 G^2}{\varepsilon^2}$ and $\gamma = \frac{D}{G\sqrt{T}}$ returns $\vec{x}$ s.t. $f(\vec{x}) \le f(\vec{x}^*) + \varepsilon$.

Proof: We'll show that on every iteration,

$$\star \quad \|\vec{x}^{(t)} - \vec{x}^*\|_2^2 \le \|\vec{x}^{(t-1)} - \vec{x}^*\|_2^2 + 2\gamma \nabla f(\vec{x}^{(t-1)})^T(\vec{x}^* - \vec{x}^{(t-1)}) + \gamma^2 \|\nabla f(\vec{x}^{(t-1)})\|_2^2.$$

$$\text{so, } f(\vec{x}^{(t-1)}) + \left(\frac{\|\vec{x}^{(t)} - \vec{x}^*\|_2^2 - \|\vec{x}^{(t-1)} - \vec{x}^*\|_2^2}{2\gamma}\right) \le f(\vec{x}^{(t-1)}) + \nabla f(\vec{x}^{(t-1)})^T(\vec{x}^* - \vec{x}^{(t-1)}) + \frac{\gamma}{2}\|\nabla f(\vec{x}^{(t-1)})\|_2^2$$

$$\Rightarrow \frac{1}{T}\sum_{t=1}^{T} f(\vec{x}^{(t-1)}) + \left(\frac{\|\vec{x}^{(t)} - \vec{x}^*\|_2^2 - \|\vec{x}^{(0)} - \vec{x}^*\|_2^2}{2\gamma T}\right) \le f(\vec{x}^*) + \frac{\gamma G^2}{2}$$

(telescoping sum...)

$$\text{convex } f \quad f\left(\frac{1}{T}\sum_{t=1}^{T} \vec{x}^{(t-1)}\right) + \frac{\|\vec{x}^{(T)} - \vec{x}^*\|_2^2 - \|\vec{x}^{(0)} - \vec{x}^*\|_2^2}{2\gamma T} \le f(\vec{x}^*) + \frac{\gamma G^2}{2}$$

(V1: ≤ 0; V1: $\le D^2$)

$$\Rightarrow f\left(\frac{1}{T}\sum_{t=0}^{T-1} \vec{x}^{(t)}\right) \le f(\vec{x}^*) + \frac{D \cdot \varepsilon \cdot G^2}{G \cdot DG \cdot 2} + \frac{G \varepsilon D^2}{2 \cdot D \cdot DG} = f(\vec{x}^*) + \varepsilon \text{ as needed.}$$

So, to prove $\star$: we know

$$\begin{aligned}\|\vec{x}^{(t)} - \vec{x}^*\|_2^2 &= \|\Pi_K(\vec{x}^{(t-1)} - \gamma \nabla f(\vec{x}^{(t-1)})) - \vec{x}^*\|_2^2 \quad \text{(by def. of } \vec{x}^{(t)}\text{)} \\ &\le \|\vec{x}^{(t-1)} - \gamma \nabla f(\vec{x}^{(t-1)}) - \vec{x}^*\|_2^2 \quad \text{(by earlier calc.)} \\ &= \|\vec{x}^{(t-1)} - \vec{x}^*\|_2^2 - 2\gamma \nabla f(\vec{x}^{(t-1)})^T(\vec{x}^{(t-1)} - \vec{x}^*) + \gamma^2 \|\nabla f(\vec{x}^{(t-1)})\|_2^2 \quad \blacksquare\end{aligned}$$

What is good about this analysis is that there is potentially no dependence on the dimension n. (If it appears, it is in the D or G terms.) This is potentially stronger than the multiplicative weights analysis (and can be applied to varying f as well), especially if our constraints are expressed in the Euclidean (ℓ_2) norm.

Note that in particular, projection to a ℓ_2-ball $B(\vec{x}_0, r)$ is easy: if $\|\vec{x}_0 - \vec{y}\|_2 \le r$, return $\vec{y}$. Else, return $\vec{x}_0 + r\frac{\vec{y} - \vec{x}_0}{\|\vec{y} - \vec{x}_0\|_2}$.

But, the running time is still pseudopolynomial: there is a factor of $T = \frac{D^2 G^2}{\varepsilon^2}$. The problem is that the step size is very small.

For a nicer objective f, we can do better.

Defn: f is β-smooth if $\forall \vec{x},\vec{y} \in K \quad \|\nabla f(\vec{x}) - \nabla f(\vec{y})\|_2 \le \beta \|\vec{x}-\vec{y}\|_2$

Lemma For β-smooth f on convex K, $|f(\vec{x}) - f(\vec{y}) - \nabla f(\vec{y})^T(\vec{x}-\vec{y})| \le \frac{\beta}{2}\|\vec{x}-\vec{y}\|_2^2$

Proof: we can write $f(\vec{x}) - f(\vec{y}) = \int_0^1 \nabla f(\vec{y} + t\cdot(\vec{x}-\vec{y}))^T(\vec{x}-\vec{y})\,dt$. Then by Cauchy-Schwarz, $|f(\vec{x}) - f(\vec{y}) - \nabla f(\vec{y})^T(\vec{x}-\vec{y})| \le \int_0^1 \|\nabla f(\vec{y}+t(\vec{x}-\vec{y})) - \nabla f(\vec{y})\|_2 \|\vec{x}-\vec{y}\|_2 dt$

By β-smoothness, $\le \int_0^1 \beta \|\vec{y} + t(\vec{x}-\vec{y}) - \vec{y}\|_2 \|\vec{x}-\vec{y}\|_2 dt$

$= \|\vec{x}-\vec{y}\|_2^2 \cdot \beta \cdot \int_0^1 |t|\,dt = \frac{\beta}{2}\|\vec{x}-\vec{y}\|_2^2$ ∎

In other words, β-smoothness gives a nice upper bound for $\vec{x}$ near $\vec{y}$: $f(\vec{x}) \le f(\vec{y}) + \nabla f(\vec{y})^T(\vec{x}-\vec{y}) + \frac{\beta}{2}\|\vec{x}-\vec{y}\|_2^2$. Complementing this,

Defn: f is α-strongly convex if $\forall \vec{x},\vec{y} \in K \quad f(\vec{x}) \ge f(\vec{y}) + \nabla f(\vec{y})^T(\vec{x}-\vec{y}) + \frac{\alpha}{2}\|\vec{x}-\vec{y}\|_2^2$

Functions that are both smooth and strongly convex thus satisfy $f(\vec{x}) = f(\vec{y}) + \nabla f(\vec{y})^T(\vec{x}-\vec{y}) + \Theta(\|\vec{x}-\vec{y}\|_2^2)$ where the hidden term lies in the interval $[\frac{\alpha}{2}, \frac{\beta}{2}]$. We'll then be able to take a much more aggressive step, $\gamma = \frac{1}{\beta}$, and obtain a running time that only grows linearly with # digits of precision (logarithmically in $\frac{1}{\epsilon}$ – "linear convergence").

Theorem For β-smooth and α-strongly convex f on closed and convex K, $\vec{x}^* = \arg\min_{\vec{x}\in K} f(\vec{x})$, projected gradient descent with $\gamma = \frac{1}{\beta}$ obtains $\|\vec{x}^{(t)} - \vec{x}^*\|_2 \le e^{-t\cdot\frac{\alpha}{\beta}}\|\vec{x}^{(0)} - \vec{x}^*\|_2^2$

Proof: First note, by our earlier lemma, since f is β-smooth,

$f(\vec{x}^{(t+1)}) - f(\vec{x}^{(t)}) \le \nabla f(\vec{x}^{(t)})^T(\vec{x}^{(t+1)} - \vec{x}^{(t)}) + \frac{\beta}{2}\|\vec{x}^{(t+1)} - \vec{x}^{(t)}\|_2^2$

where $\vec{x}^{(t+1)} = \pi_K\left(\vec{x}^{(t)} - \frac{1}{\beta}\nabla f(\vec{x}^{(t)})\right)$, and by our earlier calculation,

$\left(\vec{x}^{(t+1)} - \left(\vec{x}^{(t)} - \frac{1}{\beta}\nabla f(\vec{x}^{(t)})\right)\right)^T(\vec{x}^{(t+1)} - \vec{x}^*) \le 0$, i.e.,

$\nabla f(\vec{x}^{(t)})^T(\vec{x}^{(t+1)} - \vec{x}^*) \le -\beta(\vec{x}^{(t+1)} - \vec{x}^{(t)})^T(\vec{x}^{(t+1)} - \vec{x}^*)$. We also have by α-strong convexity that $f(\vec{x}^{(t)}) - f(\vec{x}^*) \le \nabla f(\vec{x}^{(t)})^T(\vec{x}^{(t)} - \vec{x}^*) - \frac{\alpha}{2}\|\vec{x}^{(t)} - \vec{x}^*\|_2^2$, so

$f(\vec{x}^{(t)}) - f(\vec{x}^*) + f(\vec{x}^{(t+1)}) - f(\vec{x}^{(t)}) \le \nabla f(\vec{x}^{(t)})^T(\vec{x}^{(t+1)} - \vec{x}^*) + \frac{\beta}{2}\|\vec{x}^{(t+1)} - \vec{x}^{(t)}\|_2^2 - \frac{\alpha}{2}\|\vec{x}^{(t)} - \vec{x}^*\|_2^2$

$f(\vec{x}^{(t+1)}) - f(\vec{x}^*) \le -\beta(\vec{x}^{(t+1)} - \vec{x}^{(t)})^T(\vec{x}^{(t+1)} - \vec{x}^*) + \frac{\beta}{2}\|\vec{x}^{(t+1)} - \vec{x}^{(t)}\|_2^2 - \frac{\alpha}{2}\|\vec{x}^{(t)} - \vec{x}^*\|_2^2$

$\frac{\beta}{2}\left(\|\vec{x}^{(t+1)} - \vec{x}^{(t)} - (\vec{x}^{(t+1)} - \vec{x}^*)\|_2^2 - \|\vec{x}^{(t+1)} - \vec{x}^*\|_2^2\right)$

so: $f(\vec{x}^{(t+1)}) - f(\vec{x}^*) \le \frac{\beta}{2}\left(\left(1-\frac{\alpha}{\beta}\right)\|\vec{x}^{(t)}-\vec{x}^*\|_2^2 - \|\vec{x}^{(t+1)}-\vec{x}^*\|_2^2\right)$

where $f(\vec{x}^*) \le f(\vec{x}^{(t+1)})$, so $\|\vec{x}^{(t+1)}-\vec{x}^*\|_2^2 \le \left(1-\frac{\alpha}{\beta}\right)\|\vec{x}^{(t)}-\vec{x}^*\|_2^2$.

Hence, $\|\vec{x}^{(t)}-\vec{x}^*\|_2^2 \le \left(1-\frac{\alpha}{\beta}\right)^t \|\vec{x}^{(0)}-\vec{x}^*\|_2^2 \le e^{-t\frac{\alpha}{\beta}}\|\vec{x}^{(0)}-\vec{x}^*\|_2^2$ ▨

Note that for smooth and strongly convex f, $|f(\vec{x})-f(\vec{x}^*)|$ is small iff $\|\vec{x}-\vec{x}^*\|_2$ is small: recall, $\nabla f(\vec{x}^*)^T(\vec{x}^*-\vec{x}) \le 0$.

By strong convexity, $f(\vec{x}) \ge f(\vec{x}^*) + \underbrace{\nabla f(\vec{x}^*)^T(\vec{x}-\vec{x}^*)}_{\ge 0} + \frac{\alpha}{2}\|\vec{x}-\vec{x}^*\|_2^2$

so $\frac{\alpha}{2}\|\vec{x}-\vec{x}^*\|_2^2 \le f(\vec{x}) - f(\vec{x}^*)$.

Conversely, $f(\vec{x}) - f(\vec{x}^*) \le \nabla f(\vec{x}^*)^T(\vec{x}-\vec{x}^*) + \frac{\beta}{2}\|\vec{x}-\vec{x}^*\|_2^2$

by β-smoothness $\le \frac{\beta}{2}\|\vec{x}-\vec{x}^*\|_2^2$ if $\vec{x}^*$ is in the interior (since then $\nabla f(\vec{x}^*) = \vec{0}$) or

by Cauchy-Schwarz, $\le \|\nabla f(\vec{x}^*)^T\|_2 \|\vec{x}-\vec{x}^*\|_2 + \frac{\beta}{2}\|\vec{x}-\vec{x}^*\|_2^2$