Intro & Overview, Gradient Descent.

Monday, January 6, 2020 5:42 PM

Algorithms are discrete-time procedures to solve problems.

e.g. SAT

Does f have a SATISFYING assignment?

What should my next move be?

Translate "Vágyok boldog" to English

Optimization: $\min_{x \in \Omega} f(x)$.

very general. E.g. $\exists x : f(x) = 1$

$\iff \min |f(x) - 1|$

<u>also intractable</u>. $f(x) = \begin{cases} 1 & x \neq x^* \\ 0 & \text{o.w.} \end{cases}$ $\cdot x^*$

$\min f(x)$

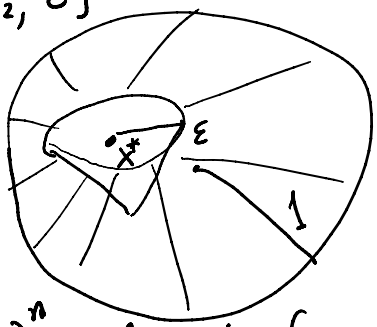
takes an infinite # evaluations of f.

But what if f is Lipshitz, i.e. $\|\nabla f\| \leq 1$.

$L(f) = \sup_{x,y} \frac{|f(x) - f(y)|}{\|x - y\|}$. and with bounded domain.

$f(x) = \min\{\|x - x^*\|_2, \varepsilon\}$ $x \in B_2^n$

is 1-Lipshitz.



$\text{Vol}(\{x : \|x - x^*\|_2 \leq \varepsilon\}) \leq \varepsilon^n \cdot \text{Vol}(B_2^n)$

Finding x takes $\Omega\left(\frac{1}{\varepsilon}\right)^n$ calls to f.

Finding x takes $\Omega\left(\frac{1}{\varepsilon}\right)^n$ calls to f.

Exercise. Show that $O\left(\frac{1}{\varepsilon}\right)^n$ calls suffice.

P2 Sampling. $f: \mathbb{R}^n \to \mathbb{R}$, sample x with density

$$\nu(x) \propto e^{-f(x)}.$$

e.g. $f(x) = \frac{\|x\|^2}{2}$, $f(x) = \begin{cases} 0 & x \in K \\ \infty & \text{o.w.} \end{cases}$

OPT → Sampling (binary search)

∴ intractable!

Examples in practice: flows, matching, LP regression, sampling Gaussian, convex body, logistic regression.

$\frac{1}{n}\sum_{i=1}^{n} f((a^T x_i)\cdot y_i) + \lambda\|a\|$

e.g. $f(z) = \log(1+e^z)$.

Suppose we assume f is convex.

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$$

then all kinds of wonderful things happen!

① Any local minimum $\nabla f(x^*) = 0 \Rightarrow$ global min $f(x^*) \le f(x)\ \forall x$.

Why? Lemma. $f(y) \ge f(x) + \nabla f(x)^T (y-x)$

Pf: $g(\lambda) = f((1-\lambda)x + \lambda y)$ $\quad g(0) = x$ $\quad g(1) = y$

by convexity $\quad g(\lambda) \le (1-\lambda) g(0) + \lambda g(1)$.

$$g(1) \ge g(0) + \frac{g(\lambda) - g(0)}{\lambda}$$

$\lambda \to 0$

$$g(1) \ge g(0) + g'(0)$$

$$g'(0) = \nabla f(x)^T (y-x)$$

$$f(y) \ge f(x) + \nabla f(x)^T (y-x).$$

closed ∴ $x, y \in K \Rightarrow [x, y] \subseteq K$.

- $f(y) \geq f(x) + \nabla f(x)(y-x)$.

② Seperation. K is closed convex, i.e. $x, y \in K \Rightarrow [x,y] \subseteq K$.

$y \notin K \Rightarrow \exists a$: $a^T y > \max_{x \in K} a^T x$.

This lets us do binary search!

Examples of convex functions.

Gradient Descent.

while $\|\nabla f(x)\| > \varepsilon$:

$x \leftarrow x - h \nabla f(x)$

Goal: Find x: $\|\nabla f(x)\| \leq \varepsilon$.

Q. how many iterations of GD? What h to use?

Why this goal: x is near-minimum in its neighborhood.

But even this is not true if ∇f can change quickly.

Assume. $\|\nabla^2 f\|_{op} \leq L$ i.e. ∇f is L-Lipshitz.

Thm. $\nabla^2 f$ is L-Lip, $\forall \varepsilon > 0$, starting at x^0, we reach x with $\|\nabla f(x)\| \leq \varepsilon$ in at most

$$2\frac{L}{\varepsilon^2}(f(x_0) - f(x^*)) \text{ steps}$$

Lem 1. $f\left(x - \frac{1}{L}\nabla f(x)\right) \leq f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$

Pf.

- $f: D \to \mathbb{R}$ $\forall x, y$

Pf:

Thm [Taylor]. $\forall$ k+1 times differentiable $g: \mathbb{R} \to \mathbb{R}$, $\forall x, y$

$$g(y) = g(x) + \sum_{i=1}^{k} \frac{(y-x)^i}{i!} g^{(i)}(x) + \frac{(y-x)^{k+1}}{(k+1)!} g^{(k+1)}(\xi) \qquad \xi \in [x, y]$$

$$g(1) = g(0) + g'(0) + \frac{1}{2} g''(z) \qquad z \in [0,1].$$

$$g(t) = f((1-\lambda)x + \lambda y)$$

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

$y = x - \frac{1}{L} \nabla f(x)$ gives

$$f\left(x - \frac{1}{L}\nabla f(x)\right) \le f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2} L \cdot \frac{\|\nabla f(x)\|^2}{L^2}$$

$$\le f(x) - \frac{1}{2L} \cdot \|\nabla f(x)\|^2.$$

Start at $f(x^0)$, min is $f(x^*)$. each step decrease is $\ge \frac{\varepsilon^2}{2L}$.

$$\Rightarrow \#\text{steps} \le \frac{2L}{\varepsilon^2} (f(x^0) - f(x^*)). \quad \blacksquare.$$

---

f convex $\qquad f(x) = x^2$

$\text{Epi}(f) = \{(x,t) : f(x) \le t\}$

Exercise: $\text{Epi}(f)$ is convex.

$$\min f(x) = \arg\min_t \text{Epi}(f).$$

min convex functions $\Leftrightarrow$ minimizing Linear functions over convex sets.

Next time: GD for convex f.

Checking convexity:

Lem: Equivalent: (1) f is convex

(2) $\forall x, y \quad f(y) \geq f(x) + \nabla f(x)^T (y-x)$

(3) $\forall x \quad \nabla^2 f(x) \succeq 0$.

Pf: (1) $\Rightarrow$ (2) earlier.

(2) $\Rightarrow$ (3) By Taylor $f(x+th) = f(x) + t\nabla f(x)^T h + \frac{t^2}{2} h^T \nabla^2 f(z) h$, $z \in [x, x+th]$

$\Rightarrow h^T \nabla^2 f(z) h \geq 0 \quad \forall h$

$t \to 0 \Rightarrow \nabla^2 f(x) \succeq 0$.

(3) $\Rightarrow$ (1) $g(\lambda) = f(\lambda x + (1-\lambda)y) - \lambda f(x) - (1-\lambda) f(y)$

$g'(\lambda) = \nabla f(\quad).(x-y) - f(x) + f(y)$

$g''(\lambda) = (x-y)^T \nabla^2 f(\quad)(x-y)$

Let $\lambda^* = \arg\max g(\lambda)$. If $\lambda^* = 0$ or 1, $\Rightarrow g(\lambda) \leq 0$ ✓

$g(1) = g(\lambda^*) + \underbrace{g'(\lambda^*)}_{=0} + \frac{1}{2} g''(z)(1-\lambda^*)^2 \quad z \in [\lambda^*, 1]$

$0 = g(1) \geq g(\lambda^*) \geq g(\lambda)$. ✓.

What about GD applied to convex f?

Th. $f(x^k) - f(x^*) \leq \frac{2LR^2}{k+4}$

$\|\nabla^2 f\|_{op} \leq L$

$\max \|x - x^0\|_2 \leq R$

$f(x) \leq f(x^0)$

$f(x) \le f(x^0)$

Pf. $\varepsilon_k = f(x^k) - f(x^*) \le \nabla f(x^k)(x^k - x^*)$

$$\le \|\nabla f(x^*)\|_2 \|x^k - x^*\|_2$$

$$\frac{\varepsilon_k}{R} \le \|\nabla f(x^*)\|.$$

Next $f(x^{k+1}) = f\left(x^k - \frac{1}{L}\nabla f(x^k)\right)$

$$\le f(x^k) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

$$\varepsilon_{k+1} \le \varepsilon_k - \frac{1}{2L}\frac{\varepsilon_k^2}{R^2}$$

$$\frac{1}{\varepsilon_{k+1}} - \frac{1}{\varepsilon_k} = \frac{\varepsilon_k - \varepsilon_{k+1}}{\varepsilon_{k+1}\varepsilon_k} \ge \frac{1}{2LR^2}.$$

Also, $\varepsilon_0 = f(x^0) - f(x^*) = \nabla f(x^*)(x^* - x^0) + \frac{1}{2}(x^* - x^0)^T \nabla^2 f(z)(x^* - x^0)$, $z \in [x^0, x^*]$

$$\le 0 + \frac{1}{2} \cdot L \cdot \|x^* - x^0\|^2$$

$$\varepsilon_0 \le \frac{1}{2}LR^2$$

$$\therefore \frac{1}{\varepsilon_k} \ge \frac{k}{2LR^2} + \frac{1}{\varepsilon_0} \ge \frac{k+4}{2LR^2}$$

$$\text{or} \quad \varepsilon_k \le \frac{2LR^2}{k+4}.$$

to get error $\propto \varepsilon$, takes $O(\frac{1}{\varepsilon})$ steps

---

How to check convexity?

- sometimes original definition
- usually $\nabla^2 f \succcurlyeq 0$.
- sometimes separation!

---

e.g. Logistic regression.

$$\min \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{(\theta^T x^i) y^i < 0\}} + \lambda \|\theta\|$$

First some examples: $\{x : f(x) \le t\}$

$x: Ax \ge b$

$x: x^T A x \le 1 \quad A \succcurlyeq 0$

$\forall v \ v^T A v \ge 0$

$\lambda_{min}(A) \ge 0$

$\det(\text{principal minors}) \ge 0$

$\|x\|_p \le 1 \quad p \ge 1$

$\|x\|_K = \{\inf t : tx \in K\}$

$0 \in K$

$\{X : X \succcurlyeq 0\}$

$\|x\|_p$

$e^x, |x|^a \ a \ge 1$

$-\log x, x \log x$

Log. Reg.

$\|x\|_p$ is convex.

$$F(a) = \sum_{i=1}^{m} f(a^T x^i)$$

$$\nabla_a F(a) = \sum_{i=1}^{m} f'(a^T x^i)\, a$$

$$\nabla^2 F(a) = \sum_{i=1}^{m} f''(a^T x^i)\, a a^T$$

So it suffices to show $f'' \geq 0$.

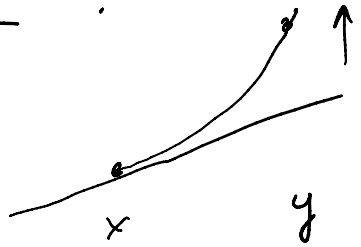
$$f(z) = \log(1+e^z)$$

$$f'(z) = \frac{e^z}{1+e^z} = 1 - \frac{1}{(1+e^z)} \qquad f''(z) = \frac{1 \cdot e^z}{(1+e^z)^2} \geq 0.$$

---

convex.

$$f(y) \geq f(x) + \nabla f(x)^T (y-x)$$



μ-Strongly convex.

$$f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2$$

Thm.

$$f(x^k) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x^0) - f(x^*)\right)$$

Pf: As before

$$f(x^{k+1}) - f(x^*) \le f(x^k) - f(x^*) - \frac{1}{2L}\|\nabla f(x^k)\|^2$$

also $$f(x^*) - f(x^k) \ge \nabla f(x^k)^T (x^* - x^k) + \frac{\mu}{2}\|x^* - x^k\|^2$$

or $$f(x^k) - f(x^*) \le \nabla f(x^k)^T (x^k - x^*) - \frac{\mu}{2}\|x^* - x^k\|^2$$

$$\arg\max_{\Delta} \nabla f(x^k)^T \Delta - \frac{\mu}{2}\|\Delta\|^2 \Rightarrow \Delta = \frac{\nabla f(x^k)}{\mu}.$$

and $$\le \frac{1}{2\mu}\|\nabla f(x^k)\|^2.$$

$$f(x^{k+1}) - f(x^*) \le f(x^k) - f(x^*)\left(1 - \frac{\mu}{L}\right).$$

Cor. After $T = \frac{L}{\mu}\log\frac{(f(x^0) - f(x^*))}{\varepsilon}$,

$$f(x^T) \le f(x^*) + \varepsilon$$

$\text{poly}\left(\frac{1}{\varepsilon}\right) \longrightarrow \log\frac{1}{\varepsilon}$.
