Gradient Descent :

- Convex set : K s.t. $\forall x, y \in K$, $\lambda x + (1-\lambda)y \in K \quad \forall \lambda \in [0,1]$.
- Convex function : $f : X \to \mathbb{R}$ is convex if

(i) $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y \in X$.

or (ii) for all lines in X, $f(x)$ restricted to the line is convex like one think of 1-dimensionality ☺.

- if f is differentiable then define the gradient of f at x :

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

such f is convex iff $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle$

↑ the linear approximation of the function f at pt x.

Eg : $f(x) = x^2$ then $\nabla f(x) = 2x$ and linear approximation is

$$y^2 \overset{?}{\geq} f(x) + \langle \nabla f(x), y - x \rangle$$
$$= x^2 + 2x(y-x)$$
$$\Rightarrow y^2 - 2xy + x^2 \overset{?}{\geq} 0$$ which is true.

If f is twice differentiable then Hessian(f) = $H(f) = \nabla^2 f$.

such f is convex iff $H(f)$ is positive semidefinite (all its eigenvalues ≥ 0).

(analog of 2nd derivative being positive).

Problems today :

$\min_{x \in \mathbb{R}^n} f(x)$ ← unconstrained minimization

$\min_{x \in K} f(x)$ ← constrained fn. minimization

(1) we'll give approximate solutions :- additive approximations.

i.e. if x^* is the minimizer then find $\hat{x}$ s.t. $f(\hat{x}) - f(x^*) \leq \varepsilon$.

(2) Need to assume something about f, and about the distance b/w x_{start} & x^*.

Simplest case [minimal assumptions] ← can we assume even less?

- f is differentiable, convex
- f is G-Lipschitz i.e. $|f(x) - f(y)| \le L\|x-y\|$ ← assume Euclidean norm, say, for now.
  $\forall x, y \in K$

$\Leftrightarrow \|\nabla f(x)\| \le G \quad \forall x \in K.$

Btw: let's figure out some properties of minimizer x^* of $f(x)$. (convex)

- unconstrained: x^* is a global minimizer of f in $\mathbb{R}^n$
  $\Leftrightarrow$ x^* is a local min of f in $\mathbb{R}^n$
  $\Leftrightarrow \nabla f(x^*) = \vec{0}$
- constrained ($x \in K$): x^* is global min $\Leftrightarrow$ x^* is local min in K
  $\Leftrightarrow \langle \nabla f(x^*), y - x^* \rangle \ge 0 \quad \forall y \in K.$

- Let's do unconstrained minimization $\min_{x \in \mathbb{R}^n} f(x)$
  (diff, convex, G-Lipschitz).

  Assume that x_0 is known, such that $\|x_0 - x^*\| \le D$ ← diameter.

Will find ϵ-approximate minimizer; i.e. $\hat{x}$ s.t. $f(\hat{x}) - f(x^*) \le \epsilon$.

in time poly $(G, D, 1/\epsilon)$.

Intuition:
$\min \frac{1}{2}\|y - x\|^2 + \eta \langle \nabla f(x), y - x \rangle$

Algorithm: Gradient Descent

Start at $\bar{x}_0 \in \mathbb{R}^n$

At each timestep $\bar{x}_{t+1} \leftarrow \bar{x}_t - \eta_t \cdot \nabla f(\bar{x}_t)$

Repeat for T steps.

Q: What is η_t? For this algo, we'll set $\eta_t = \eta$ for some η.

Q: What is T?

And what guarantees can we get on final fn value?

Thm 1: Given ϵ, x_0 s.t $\|x^0 - x^*\| \le D$, function f diff, convex, G-Lipschitz,

the gradient descent algorithm produces $x_1, x_2 \dots x_T$

such that $f(x_T) - f(x^*) \le \epsilon$ if $T \le \left(\frac{GD}{\epsilon}\right)^2$.

[N.b. if G, D are constants, then this is like $T = O(1/\epsilon^2)$.]

[There are examples s.t this guarantee is tight].

Pf: $f(x_t) - f(x^*) \le \langle \nabla f(x_t), x_t - x^* \rangle$

$= \langle \frac{1}{\eta}(x_t - x_{t+1}), x_t - x^* \rangle$

$= \frac{1}{\eta} \cdot \frac{1}{2}\left(\|x_t - x_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$

$= \frac{1}{2\eta}\left(\|\eta \nabla f(x_t)\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$

$\|\eta \nabla f(x_t)\|^2 \le \eta^2 G^2$; $\|x_t - x^*\|^2$ and $\|x_{t+1} - x^*\|^2$ will telescope

$\langle \bar{a}, \bar{b} \rangle = \left(\|\bar{a}\|^2 + \|\bar{b}\|^2 - \|\bar{a} - \bar{b}\|^2\right)^{\frac{1}{2}}$

$\Rightarrow \frac{1}{T}\sum_t^T (f(x_t) - f(x^*)) \le \frac{1}{2\eta}\left[\eta^2 G^2 + \frac{\|x_0 - x^*\|^2}{T}\right]$ drop negative term.

$= \frac{1}{2\eta}\left[\eta^2 G^2 + \frac{D^2}{T}\right]$.

[apply Jensen's, and define $\hat{x} = \frac{1}{T}\sum x_t$

$f(\hat{x}) = f\left(\frac{1}{T}\sum_t^T x_t\right) \le \frac{1}{T}\sum f(x_t)$.]

$\Rightarrow f(\hat{x}) - f(x^*) \le \frac{1}{2}\left[\eta G^2 + \frac{D^2}{\eta T}\right]$

Set $\eta = \frac{D}{G\sqrt{T}}$ gives $\le \frac{DG}{\sqrt{T}} \le \epsilon$

If we set $T \ge \left(\frac{DG}{\epsilon}\right)^2$.

Btw: if we set $\eta_t = \frac{D}{\sqrt{t} \cdot G}$ then lose a small $\log T$ term but now not dep. on length.

Not a polytime algo, since rate of convergence is slow.

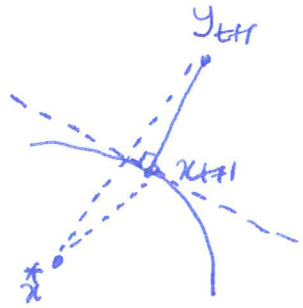
Basically get one more bit of info ($\epsilon \to \epsilon/2$) need 4 times more time.

$\Rightarrow$ for n bits will need exponential time. Want better.

But ∃ tight examples if we don't assume wave ← (soon).

But it's a great start :-

① Constrained optimization $\min_{x \in K} f(x)$.



Same guarantees. Almost same algorithm; change thus:-

$$y_{t+1} \leftarrow x_t - \eta \nabla f(x_t).$$

but y_{t+1} may be outside K.

so def $x_{t+1} \leftarrow \Pi_K y_{t+1}$ = nearest point in K to y_{t+1}
= "projection" of y_{t+1} onto K.

[Lemma: $\|y_{t+1} - x^*\|^2 \geq \|x_{t+1} - x^*\|^2$

B/c the angle between 

must be obtuse, since $x^*, x_{t+1} \in K$ and K is convex.

⇒ do the same analysis:-

$$f(x_t) - f(x^*) \leq \cdots \leq \frac{1}{2\eta}\left[\eta^2 G^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2\right]$$
$$\leq \frac{1}{2\eta}\left[\eta^2 G^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right] \cdots$$
and same follows.

⇒ no change.

Can do other things: eg [Frank Wolfe algorithm]: find the optimizer in K in the direction of $\nabla f(x_t)$, say $z_{t+1} = \min_{z \in K} \langle \nabla f(x_t), z \rangle$ ← linear optimization
and take a small step in the direction of z_{t+1}. to get x_{t+1}.

Guarantees?

(2) Online convex optimization:

Each time algorithm picks point x_t

Then adversary picks function (convex) $f_t(\,)$.

sps $\|\nabla f_t(x)\| \le G$ $\forall t$ $\forall x \in K$.

Cost $= f_t(x_t)$. want to minimize regret.

Algo: same. just use $x_{t+1} \leftarrow x_t - \eta \nabla f_t(x_t)$.

Analysis: shows $\frac{1}{T}\left(\sum_t f_t(x_t) - f_t(x^*)\right) \le \frac{DG}{\sqrt{T}} \le \epsilon$ if $T \ge \frac{(DG)^2}{\epsilon^2}$.

- Same $\frac{1}{\epsilon^2}$ dependence, but now depends on D, G not on $\lg N$ (↑ #experts).

- However, holds for all convex bodies K, not just for the probability simplex.

What does it give for $K = \Delta_N$? and say loss functions $l_t(\vec{x}) = \sum_i l_i^t x_i$

$\|\nabla l_t\|_2 \le \|\mathbb{1}\|_2 \le \sqrt{N}$ since each $l_i^t \in [-1, 1]$.

$\text{diam}(K) = 1$.

$\Rightarrow$ need $T \ge \frac{N}{\epsilon^2}$ for ϵ-regret-on-average.

much worse. (but we'll do better).

---

(3) What if not differentiable?

Use "subgradients". need a vector $\nabla f(x_t)$ such that

$$f(y) \ge f(x_t) + (\nabla f(x_t), y - x_t).$$

If not diff, may have many such vectors. [Set called $\partial f(x_t)$]

Take any of these. [also if approximate subgradients, can lose a little δ and still be OK etc.].

N.b. • no dependence on the dimension n. [not explicitly, anyways].

• Could choose a single coordinate to move along — at random

$$x_{t+1} \leftarrow x_t - \eta_t \langle \nabla f(x_t), e_i \rangle \cdot e_i \quad \text{[Coordinate descent]}$$

Naturally slows down things by factor of $n \leftarrow$ dimension.

How do you get the gradient?

- Compute it explicitly
- estimate it explicitly.
- or "stochastic gradient" descent:- $\nabla f(x) = \lim_{\delta \to 0} \mathop{\mathbb{E}}_{u \in B(0,\delta)} \left[\frac{f(x+u) - f(x)}{\delta} \cdot u \right]$

So can use a single sample to move along.

Q: how noisy is this estimator ← depends, see lots of work on this

---

But now: if we assume more, we get more.

- Def: f is ℓ-strongly convex if (locally "grows at least as fast as a quadratic")

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\ell}{2} \|y - x\|^2.$$

Fact: f is ℓ-strongly convex $\iff$ $H(f) \succeq \ell I$. (min eigenvalue $\geq \ell$).

For this case: almost same analysis works.

choose a step size $\eta_t \propto \frac{1}{\ell t}$ or so (constants dep on G, D dropped).

get that $T = \frac{O(G^2 \lg T)}{\ell \epsilon}$ is enough [HW]

Even faster?

Def: f is β-smooth if ("locally grows no faster than a quadratic")

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\beta}{2} \|y - x\|^2.$$

Also: equivalent to $\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$

$\Rightarrow$ gradients are also β-Lipschitz.

Fact: f is β-smooth $\iff$ $H(f) \preceq \beta I$.

For this case too, slightly different analysis shows that $T = \Theta(\frac{1}{\epsilon})$ enough

---

What if we assume both? $\ell I \preceq H(f) \preceq \beta I$

f is "well conditioned"

Then gradient descent gives a $\lg(\frac{1}{\epsilon})$ convergence. !!

Also called "linear convergence" ← one bit of accuracy per constant # of steps.

Assume that f is both β-smooth and ℓ-strongly-convex $\quad \ell I \preceq \nabla^2 f(x) \preceq \beta I$.

Fact 1: $f(x_t) - f(x^*) \le \dfrac{\|\nabla f(x_t)\|^2}{2\ell}$.

Pf: $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\ell}{2}\|y - x\|^2$ by ℓ-strong-convexity

$$\ge \min_z \left\{ f(x) + \langle \nabla f(x), z - x \rangle + \frac{\ell}{2}\|z - x\|^2 \right\}$$

$$= f(x) - \frac{1}{2\ell}\|\nabla f(x)\|^2$$

$$\Rightarrow f(x_t) - f(x^*) \le \frac{1}{2\ell}\|\nabla f(x_t)\|^2$$

$\blacksquare$

$\nabla f(x) = \ell \cdot (x - z)$

$\Rightarrow z = x - \frac{1}{\ell}\nabla f(x)$

plug back in

Claim: $f(x_{t+1}) - f(x^*) \le \exp\left(-\frac{t}{(\beta/\ell)}\right)(f(x_0) - f(x^*))$

(β/ℓ ↑ condition #)

Pf: let $\Delta_t = f(x_t) - f(x^*)$.

$$\Rightarrow \Delta_t - \Delta_{t-1} = f(x_t) - f(x_{t-1})$$

$$\le \langle \nabla f(x_{t-1}), x_t - x_{t-1} \rangle + \frac{\beta}{2}\|x_t - x_{t-1}\|^2$$

$$= -\eta_t \|\nabla f(x_{t-1})\|^2 + \frac{\beta}{2}\eta_t^2 \|\nabla f(x_{t-1})\|^2$$

set $\eta_t = 1/\beta$

$$= -\frac{1}{2\beta}\|\nabla f(x_{t-1})\|^2 \le -\frac{1}{2\beta} \cdot 2\ell\,\Delta_{t-1}$$

from fact 1.

$$= -(\ell/\beta)\,\Delta_{t-1}$$

$$\Rightarrow \Delta_t \le (1 - \ell/\beta)\,\Delta_{t-1}$$

$$\le \Delta_0 e^{-(\ell/\beta)t}$$

$\blacksquare$

Can also show convergence $x_t \to x^*$ (see the other proof)

Theorem: Suppose f is both β-smooth & ℓ-strongly convex.

Then if $x^* \leftarrow \min_{x \in \mathbb{R}^n} f(x)$ [unconstrained] $\qquad \ell I \preceq H(f) \preceq \beta I$.

we get ① $\|x_t - x^*\|^2 \leq e^{-\delta t} \|x_0 - x^*\|^2$ for some $\delta = \delta(\beta, \ell)$

② $f(x_t) - f(x^*) \leq \frac{\beta}{2} \|x_0 - x^*\|^2 \exp\left(-\frac{4t}{\beta/\ell}\right)$

(β/ℓ = condition # of the function.

if we set $\eta = \frac{2}{\ell + \beta}$

Proof: [will skip in class].