Gradient Descent

- convex fn. basic defns; defn 1st order, $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle$
  2nd order: $H_f(x) \succeq 0$.

- unconstrained min / constrained min $\qquad K$ = convex set $\{\lambda y + (1-\lambda)z \in K \; \forall y,z \in K \; \forall \lambda \in [0,1]\}$.

(*) f differentiable, convex, Lipschitz $\rightarrow$ $|f(x)-f(y)| \leq G \cdot \|x-y\| \; \forall x,y \in K$ or $\|\nabla f(x)\| \leq G \; \forall x \in K$

can extend some of this to the nondifferentiable case via subgradients.

Unconstrained: $f(x^*)$ is minimizer if $\nabla f(x^*) = 0$. $\qquad$ local = global

Constrained: $f(x^*)$ is minimizer if $\forall y$, $\langle \nabla f(x^*), y - x^* \rangle \geq 0$. $\qquad$ "gradient in same dir as all other pts".

Pf $f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x^* \rangle \geq f(x^*)$. by above defns.

* So if $\exists$ closed form may be able to find minimizer by differentiation. But not if unconstrained, or no clean closed form.

Assume we have some x_0, $\|x^* - x_0\| \leq D$ "diameter".

Want to find $\hat{x}$ st $f(\hat{x}) - f(x^*) \leq \varepsilon$. & fast.

$$\boxed{\text{Claim: if \#steps } T \geq \left(\frac{GD}{\varepsilon}\right)^2 \text{ then } f(\hat{x}) - f(x^*) \leq \varepsilon}$$

Here $\eta_t = \eta = D/G\sqrt{T}$

Algo: Gradient descent (x_0)

- at each time $t+1 = 1, 2, \ldots T$

  $x_{t+1} \leftarrow x_t - \eta_t \cdot \nabla f(x_t)$

  Output $\hat{x} = \frac{1}{T} \sum x_t$.

Pf: Let's look at how

$f(x_t) - f(x^*)$ behaves over time t.

Define $\Phi_t = \frac{1}{2\eta} \| x_t - x^* \|^2$.

Suppose $f(x_t) - f(x^*) + \Delta\Phi_t \leq$ blah. $\quad \forall t$ $\qquad (\Phi_{t+1} - \Phi_t)$

$$\Rightarrow \sum_t (f(x_t) - f(x^*)) \leq T.\text{blah} + (\Phi_{init} - \Phi_{final})$$

$$\Rightarrow \underset{f(\frac{1}{T}\sum x_t)}{f(\hat{x})} - f(x^*) \leq \frac{1}{T} \sum_t (f(x_t) - f(x^*)) \leq \underbrace{\text{blah} + \frac{1}{T}\Phi_{init}}_{\text{algebra to show } \leq \varepsilon}$$

$$f(x_t) - f(x^*) + \Delta\Phi_t \longrightarrow \frac{1}{2\eta}\left[\|x_{t+1} - x^*\|^2 - \|x_t - x^*\|^2\right]$$

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

I $\begin{cases} \|a+b\|^2 \\ = \|a\|^2 + \|b\|^2 \\ \quad + 2\langle a, b\rangle \end{cases}$

II $\begin{cases} x_{t+1} - x^* \\ = (x_{t+1} - x_t) = -\eta\nabla f(x_t) \\ \quad + (x_t - x^*). \end{cases}$

by defn of GD.

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

$$= \frac{\eta}{2}\nabla_t^2 - \langle \nabla f(x_t), x_t - x^*\rangle.$$

$$= f(x_t) - f(x^*) - \langle \nabla f(x_t), x_t - x^*\rangle + \frac{\eta}{2}\nabla_t^2.$$

$$\le \frac{\eta}{2}\nabla_t^2 \le \frac{\eta}{2}G^2.$$

$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle$

$\Rightarrow$ get $f(\hat{x}) - f(x^*) \le \frac{\eta}{2}G^2 + \frac{1}{2\eta}D^2 \cdot \frac{1}{T}$ set $\eta = \frac{D}{G\sqrt{T}}$ to get

$$\frac{DG}{2\sqrt{T}} + \frac{DG}{2\sqrt{T}} = \frac{DG}{\sqrt{T}} \le \varepsilon \quad \text{if} \quad T \ge \left(\frac{DG}{\varepsilon}\right)^2.$$

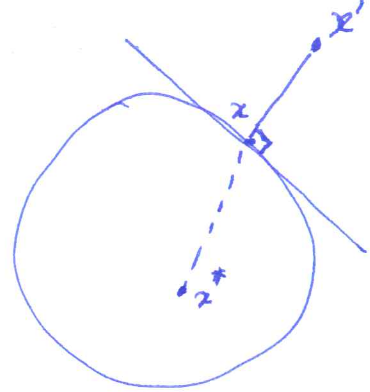
So: if you want to be within ε of OPT, need $1/\varepsilon^2$ time. Slow. For 1 more bit of info, need to run for 4 times more step.

But $\exists$ tight examples that show basic GD (without any further assumptions) does $1/\varepsilon^2$. ☹

However: constrained opt: take step and then project.

$$x'_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$$

$$x_{t+1} \leftarrow \Pi_K(x'_{t+1}). \quad \text{closest pt.}$$



Claim: $\|x_{t+1} - x^*\|^2 \le \|x'_{t+1} - x^*\|^2$.

Pf: Pythagoras.

Also: Frank Wolfe algorithm.

Online Convex Opt: ☺

[Thm: Fix $\varepsilon > 0$. ~~Given~~ Want algo s.t. we play $x_1 x_2 \dots x_T$
world plays $f_1 f_2 \dots f_T$ convex fns. then $\forall x^* \in K$.

Avg Regret $= \frac{1}{T}\sum f_t(x_t) \le \frac{1}{T}\sum f_t(x^*) - \varepsilon$ after $T \geq \frac{\text{blah}}{\varepsilon^2}$ steps

Algo (Proof): just put f_t everywhere instead of f.

Hence: get a low regret algorithm with default parameters.

For old case: $K = \Delta_N \Rightarrow D = 1$.
$G = \max_x \|\nabla f(x)\| \le \sqrt{N}$.

$\Rightarrow$ regret $\le \varepsilon$ after $T \geq \frac{N}{\varepsilon^2}$ steps, much weaker. Do better later.

Other Observations: . no explicit dependence on n (dimension "free").

- Non diff? use subgradients.
- Compute gradient? use stochastic gradients. $\nabla f(x) = \lim_{\delta \to 0} E\left[\frac{f(x+\delta \bar{u})}{\delta} \cdot \bar{u} \cdot d\right]$ one point estimate

$\bar{u}$ over ball of radius 1. (d: dimension)

Intuition: $\frac{df}{dx} \approx \frac{f(x+\delta) - f(x-\delta)}{2\delta} = E\left[\frac{f(x+\delta u)}{\delta} \cdot \bar{u} \cdot 1\right]$

in fact equality in the limit

Very noisy / high variance. So use carefully. Much work, maybe do later in course

---

Some more assumptions to speed things up!

① α-strongly convex fn: $f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \frac{\alpha}{2}\|y-x\|^2$. $\Leftrightarrow H(x) \succeq \alpha I$.

[Exercise:] With a slightly different proof (almost same) get $T = \tilde{O}(\frac{1}{\varepsilon})$ instead.
and $\eta_t \sim \frac{1}{t}$ [also in online] give intuition!!

② β-smooth: $f(x) \le$ ________ $+ \frac{\beta}{2}\|y-x\|^2$. $\Leftarrow H(x) \preceq \beta I$.

Then get $T = \tilde{O}(1/\varepsilon)$ [but not online convergence]

And suppose:

f is "well conditioned": both β-smooth & α-strongly convex

$$\alpha I \preceq H(f) \preceq \beta I$$

Then $T = O(\log 1/\varepsilon)$ suffices! [Linear convergence].

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

$$= -\eta\|\nabla_t\|^2 + \frac{\beta}{2}\|\nabla_t\|^2 \cdot \eta^2$$

Set $\eta = \frac{1}{\beta} \Rightarrow \; = -\frac{1}{2\beta}\|\nabla_t\|^2$ (*)

$f(x_t) - f(x^*) \le \langle \nabla f(x_t), x_t - x^* \rangle - \frac{\alpha}{2}\|x_t - x^*\|^2$ α-strong.

$$\le \|\nabla_t\| \cdot \|x_t - x^*\| - \frac{\alpha}{2}\|x_t - x^*\|^2.$$

$\le \frac{1}{2\alpha}\|\nabla_t\|^2$. (**)

$g \cdot d - \frac{\alpha}{2} d^2$
max wrt d $g = \alpha d$
$\Rightarrow d = g/\alpha$

$\Rightarrow f(x_{t+1}) - f(x_t) \le -\frac{1}{2\beta}\nabla^2 \le -\alpha/\beta\,(f(x_t) - f(x^*))$ by (*) and (**).

if $\Delta_t = f(x_t) - f(x^*)$

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

$$\Rightarrow \Delta_T \le e^{-\alpha/\beta\, T} \cdot \Delta_0.$$

$\Rightarrow$ to be within ε of $f(x^*)$ need $\frac{\beta}{\alpha}\log\left(\frac{\Delta_0}{\varepsilon}\right)$ rounds.

"Linear convergence"

# of rounds increases as $O(1)$ for each extra bit of accuracy.