# THEORY OF MACHINE LEARNING

# LECTURE 13

GRADIENT DESCENT, THOUGHTS

# RECAP: CONVEX OPTIMIZATION

$f$ : convex .

$\mathcal{D}$ .

- **Problem.** Given a convex function defined over a convex domain, find the minimizer (or min value).

$\underset{x \in \mathcal{D}}{\operatorname{argmin}} \; f(x)$ .

- Gradient descent – inspired by Taylor approximation
  - Start with some feasible $w_0$
  - For t = 0, 1, …, T-1, set $w_{t+1} = w_t - \eta \nabla f(w_t)$

$w_t$

$f(w_{t+1}) = f(w_t) - \eta \|\nabla f(w_t)\|^2$ .

- How do you set/"tune" the learning rate?
- Staying feasible

keep projecting to $\mathcal{D}$ .

## BASIC THEOREM

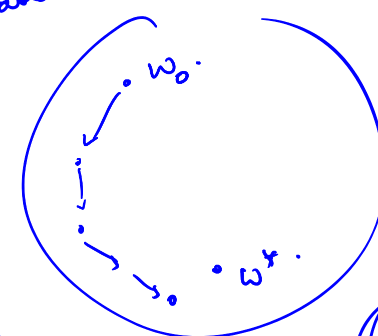
- Assume f is L Lipschitz, domain is all of $R^d$, $|w_0 - w^*| \le B$
- **Theorem.** Consider running T steps of gradient descent with a fixed learning rate $\eta$. Then we have

norm of gradient.

$\bar{w} = \frac{w_1 + w_2 + \dots + w_T}{T}$.

$$f(\bar{w}) - f(w^*) \le \frac{1}{T}\sum_{t=1}^{T} \underbrace{f(w_t) - f(w^*)}_{\ge 0.} \le \frac{B^2}{2\eta T} + \frac{L^2\eta}{2}$$

RHS becomes $\frac{BL}{\sqrt{T}}$



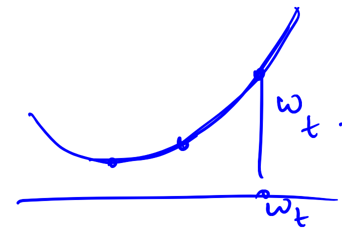

- Same proof works if we had a constrained domain
- Proof works even if functions at different time steps were different!

$T = \frac{(BL)^2}{\varepsilon^2}$

$$\sum_{t=1}^{T} f_t(w_t) - f_t(w) \le \frac{B^2}{2\eta} + \frac{L^2\eta T}{2} \qquad \forall w.$$

$\sum_t f_t(w_t) - f_t(w)$

# ANALYSIS




- Use "basic inequality" about convex functions, for any t,

$$f(w^*) \geq f(w_t) + \langle w^* - w_t, \nabla f(w_t) \rangle$$

$f(w) \geqslant f(w_t) + \langle \nabla f(w_t), w - w_t \rangle$

- Use the potential function $\boxed{\Phi_t = |w_t - w^*|^2}$
- Note that $\Phi_t - \Phi_{t+1}$ (potential drop) is lower bounded by how far $f(w_t)$ is from $f(w^*)$
- $\Phi_t - \Phi_{t+1} \geq 2\eta \left( f(w_t) - f(w^*) \right) - \eta^2 |\nabla f(w_t)|^2$ $\rightarrow \eta^2 L^2$
- Summing over t gives the bound
- Applications to online convex optimization, SGD $\rightarrow$ stochastic grad descent.

Cool thing about the analysis: can even have different $f_t$'s at diff times t.

# STOCHASTIC GRADIENT DESCENT

- Consider the setting where the function f can be decomposed as

$$f(w) = \frac{1}{m}\sum_{i=1}^{m} g_i(w)$$

L-Lipschitz.

[ true in usual ML applns, where $g_i$ correspond to one (mini) batch ]

- In iteration t, pick index $i_t$ uniformly at random and take a gradient step, i.e., $w_{t+1} = w_t - \eta \nabla g_{i_t}(w_t)$

not the same as $w_t - \eta \nabla f(w_t)$

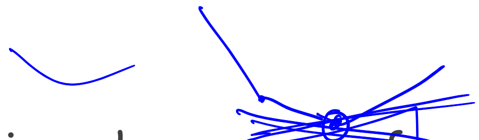
- Now $w_t$ is a random variable, and we need to argue about $E[\,f(w_t)]$
- We still have that the expected potential drop at step t (conditioned on any trajectory so far) is $2\eta\big(f(w_t) - f(w^*)\big) - \eta^2 L^2$

$\mathbb{E}\left[\Phi_t - \Phi_{t+1} \mid w_t\right]$

- Earlier bound holds <u>in expectation</u>

# NOISY GRADIENT DESCENT

$w \rightarrow w - \eta \nabla f(w)$



- Consider the setting where we perform gradient descent on the function f using a "noisy gradient oracle"
- Given a point w, suppose we get "noisy gradient"

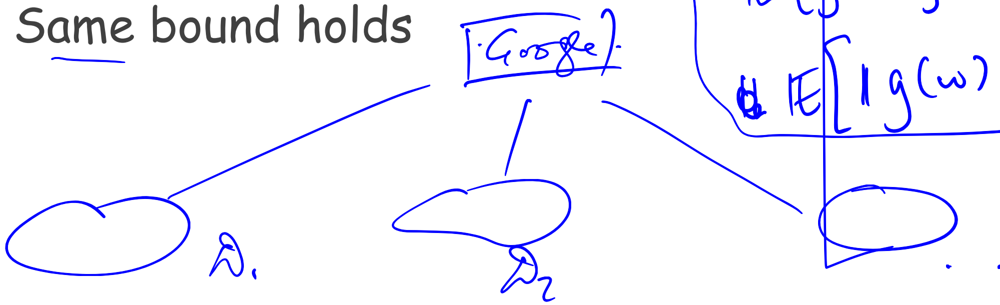
Oracle that gives $g(w) \rightsquigarrow$ noisy gradient; with the props that

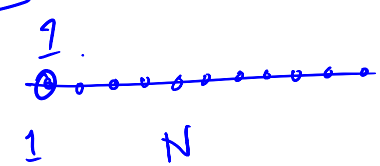
$$\mathbb{E}[g(w)] = \nabla f(w).$$

$$\mathbb{E}[\|g(w)\|^2] \leq L^2.$$

$\sqrt{N}$.

- Same bound holds



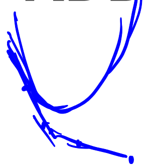



f is said to be

$ax^2+bx+c$ $\quad -\frac{b}{2a}$

$f(y) \le f(x) + \langle \nabla f(x), y-x \rangle + M\|y-x\|^2$

# ADDITIONAL STRUCTURE ON FUNCTIONS

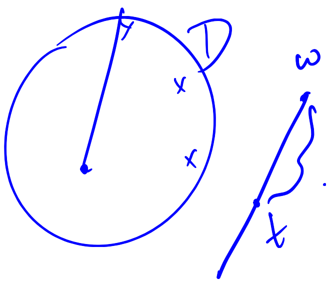


(earlier: $|f(x)-f(y)| \le L\|x-y\|$ ; $\|\nabla f(x) - \nabla f(y)\| \le M\|x-y\|$

$\|\nabla f(x)\| \le L$ $\qquad \|\nabla^2 f(x)\|_2 \le M$

- **Smoothness** – function f is M smooth if *gradient* is M-Lipschitz
- **Key observation:** in this case, every iteration yields drop in function value!

$w_{t+1} = w_t - \eta \nabla f(w_t)$ $\qquad -\eta \nabla f(w_t)$




$$f(w_{t+1}) \le f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + M\|w_{t+1} - w_t\|^2$$

$$= f(w_t) - \eta\|\nabla f(w_t)\|^2 + M\eta^2\|\nabla f(w_t)\|^2$$

$$= f(w_t) - \frac{1}{2M}G^2 + \frac{1}{4M}G^2 = f(w_t) - \frac{\eta}{2}\|\nabla f(w_t)\|^2$$

$\eta \approx \frac{1}{2M}$

- After T steps, $\sum_t |\nabla f(w_t)|^2$ is bounded by $4M\,(f(w_0) - f(w^*))$
- Convergence rate of 1/T

$$\boxed{f(w_{t+1}) \le f(w_t) - \frac{\eta}{2}\|\nabla f(w_t)\|^2}$$

$$\Leftrightarrow \|\nabla f(w_t)\|^2 \le \frac{2}{\eta}\left(f(w_t) - f(w_{t+1})\right)$$

Earlier analysis: $\Phi_t = \| w_t - w^* \|^2$

Saw that: $\Phi_t - \Phi_{t+1} \geq 2\eta\left( f(w_t) - f(w^*) \right) - \eta^2 \| \nabla f(w_t) \|^2$

$$\iff \quad f(w_t) - f(w^*) \leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \frac{\eta}{2} \| \nabla f(w_t) \|^2$$

$$\leq \frac{\Phi_t - \Phi_{t+1}}{2\eta} + \left( f(w_t) - f(w_{t+1}) \right)$$

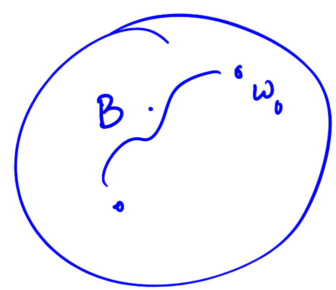
$\eta = 1/2M$

$$\frac{1}{T} \cdot \sum_{t=1}^{T} f(w_t) - f(w^*) \leq \frac{2M(\Phi_0 - \dots)}{T} + \frac{f(w_0) - f(w_T)}{T}$$

$f(w_0) - f(w_T) \leq C$

error after T steps $\simeq \frac{1}{T}$.

$$\frac{2MB^2}{T} + \frac{C}{T}.$$



# NONCONVEX (SMOOTH) FUNCTIONS

→ If $f$ is non-convex but $M$-smooth, then anysis before (with $\eta = \frac{1}{2M}$) implies that $\|\nabla f(w_t)\|^2 \le 4M(f(w_t) - f(w_{t+1}))$

$$\Rightarrow \frac{1}{T} \sum_{t=1}^{T} \|\nabla f(w_t)\|^2 \le \frac{4M(f(w_0) - f(w^*))}{T}$$

→

$$\exists t \text{ s.t. } \|\nabla f(w_t)\|^2 \le \frac{4MC}{T}.$$

↓

approximate singular pt.

# ADDITIONAL STRUCTURE ON FUNCTIONS

- **Smoothness**: function is M smooth if *gradient* is M-Lipschitz
- **Strongly convex**: function is m-strongly convex if we have a "lower bound" via a parabola

# IMPROVEMENTS, GENERALIZATIONS

- **Polyak-Lojasiewicz inequality**: suppose f satisfies:

  $|\nabla f(w)|^2 \geq c(f(w) - f(w^*))$ for all w
- "Global" condition, but can be satisfied for non-convex f

- Polyak's "heavy ball" method (momentum)
- AdaGrad and related methods
- Second order (Newton) methods
- …