THEORY OF MACHINE LEARNING

LECTURE 14

GRADIENT DESCENT – SMOOTH, STRONGLY CONVEX

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 η. Then we have

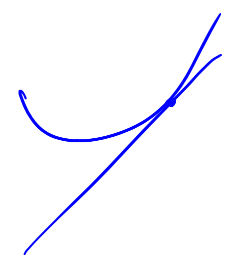
$$\frac{1}{T}\sum_{t=1}^{T} f(w_t) - f(w) \le \frac{B^2}{2\eta T} + \frac{L^2\eta}{2}$$

tuning $\rightarrow$ η.

$$\frac{LB}{\sqrt{T}}$$

- Same proof works if we had a constrained domain
- Use "basic inequality" about convex functions, for any t,

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



- Use the potential function $\Phi_t = |w_t - w^*|^2$

NOISY GRADIENT DESCENT (GENERALIZES SGD)

g.

- Doing gradient descent on f using a "noisy gradient oracle"

w

- Given a point w, suppose we get "noisy gradient"

$g(w)$

$$\boxed{\mathbb{E}[g(w)] = \nabla f(w)} \;; \quad \mathbb{E}[\|g(w)\|^2] \le L^2.$$

- Same bound holds assuming noise is unbiased, and has low variance

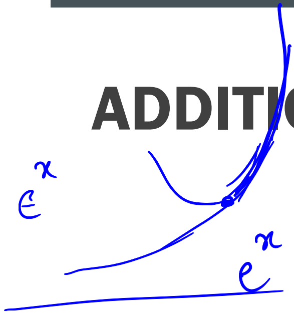
$-MI \preceq \nabla^2 f(x) \preceq MI$

psd dominated

-10, 5

magnitude of the largest eigenvalue in mag.

ADDITIONAL STRUCTURE: SMOOTHNESS



$$\|\nabla f(x) - \nabla f(y)\| \le M \cdot \|x-y\| \iff \|\nabla^2 f(x)\|_2 \le M.$$

$\forall$ unit vects z, $z^T \nabla^2 f(x) z \le M$

- **Smoothness** – function is M smooth if *gradient* is M-Lipschitz
- **Key observation:** in this case, every iteration yields drop in function value (first order approx. is accurate in ball of radius < 1/2M)

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

$$\left(\frac{\partial^2 f}{\partial x_i \partial x_j}\right)_{d\times d}$$

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

$\varepsilon = 10^{-4}$

- GD on smooth non-convex functions converges to "approximately singular" points

Matrix basics: $A \in \mathbb{R}^{d \times d}$

$z = (z_1, \ldots, z_d)$

Quadratic form in d variables $z_1, z_2, \ldots, z_d$.

$$z^T A z = \sum_{i,j} A_{ij} (z_i z_j)$$

$$\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix}$$

$$\tilde{A} = \frac{A + A^T}{2}$$

$z_1^2 - z_1 z_2 + z_2^2$

If we look at max over all z (vectors) with $\|z\| = 1$ of this quadratic form:

$Az = \lambda z$.

$z^T A z = \lambda \underbrace{(z^T z)}_{=1}$

$$\underbrace{\max_{\|z\|=1} |z^T A z|}_{\downarrow \text{ spectral norm.}}$$

$\max_{\|z\|=1} z^T A z \longrightarrow \lambda_{max}$ of the matrix.

$\min_{\|z\|=1} z^T A z$. λ_{min} (could be -ve).

$(A - B)$ is psd.

$A \succeq B$ if $\forall z$

$z^T A z \geq z^T B z$.

$\succeq$ (succeq)

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

(alt. defn of smoothness) $f(w_{t+1}) \le f(w_t) + \langle \nabla f(w_t), \underbrace{w_{t+1} - w_t}_{= -\eta \nabla f(w_t)} \rangle + M \| w_{t+1} - w_t \|^2$

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

$\eta > M\eta^2$ or $\eta < \frac{1}{M}$;

Say we set $\eta \le \frac{1}{2M}$.

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

$1/T$

CAN WE GO BEYOND 1/T CONVERGENCE?

- **Smoothness:** function is M smooth if *gradient* is M-Lipschitz

Nesterov acceleration.

- Purely assuming smoothness, can get rate of $\boxed{1/T^2}$ (Nesterov 1983)

$w_t \rightarrow w_{t+1} \rightarrow \ldots$

w'_t $\quad w_{t+1} = w_t - \eta \nabla f(w_t) - \ldots w'_t$

[Optimal for all "gradient based" methods] → Oracle l.b.s.

* · Formally, consider "GD-like" procedures, where $w_{t+1} = \mathcal{H}\big(w_1, w_2, \ldots, w_t, \nabla f(w_1), \nabla f(w_2), \ldots \nabla f(w_t)\big)$.

~~Among~~ For all procedures of this kind, error after t iterations must be $\geq \frac{1}{t^2}$ in the worst case.

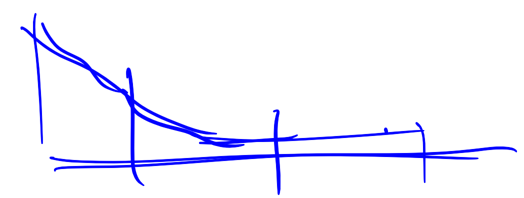
STRONG CONVEXITY

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

$$(\mu\text{-SC}): \; f(y) \geq f(x) + \langle \nabla f(x), y-x \rangle + \left(\mu \cdot \|y-x\|^2\right).$$

$$(M\text{-smooth}): \; f(y) \leq f(x) + \langle \nabla f(x), y-x \rangle + M \|y-x\|^2.$$

$$\boxed{\forall x: \quad \mu I \preceq \nabla^2 f(x) \preceq M I}$$



$$\rightarrow \; f(w^*) \geq f(w) + \underbrace{\langle \nabla f(w), w^*-w \rangle} + \mu \underbrace{\|w^*-w\|^2}_{\Phi_t}.$$

$$\Phi_t = \|w_t - w^*\|^2.$$

$$\Phi_{t+1} = \|w_t - \eta \nabla f(w_t) - w^*\|^2$$

$$= \Phi_t - \eta \langle \nabla f(w_t), w_t - w^* \rangle + \eta^2 \|\nabla f(w_t)\|^2.$$

Smoothness: $f(w_{t+1}) \leq f(w_t) - \frac{\eta}{2} \cdot \|\nabla f(w_t)\|^2$, if $\eta < \frac{1}{2M}$.

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

---

using, μ. S.C

$$\Phi_{t+1} \leq \Phi_t - \eta [f(w_t) - f(w^*)] - \eta \mu \Phi_t + \frac{2}{\eta} \left(f(w_t) - f(w_{t+1}) \right)$$

$$\underline{\underline{\Phi_{t+1}}} \leq \underbrace{\left(1-\frac{\eta\mu}{4}\right)}\cdot\Phi_t$$

$$\leq \left(1-\frac{\eta\mu}{4}\right)^2\cdot\Phi_{t-1} \qquad \eta\sim\frac{1}{2M}.$$

$$\vdots$$

$$\leq \left(1-\frac{\eta\mu}{4}\right)^{t+1}\Phi_0.$$

$z<1$

$\mu \leq M.$

$1-z \simeq e^{-z}$

After T steps, $\Phi_T \leq \left(1-\frac{\mu}{8M}\right)^T\cdot B^2 \approx e^{-\frac{\mu T}{8M}}\cdot B^2.$

if we want this to be $<\varepsilon$, then we must pick.

$$T\simeq \log\left(\frac{B^2}{\varepsilon}\right)\cdot\frac{8M}{\mu}$$

condition #

$$\underline{\underline{T\simeq\left(\frac{M}{\mu}\right)\cdot\log\left(\frac{1}{\varepsilon}\right)}}$$

IMPROVEMENTS, GENERALIZATIONS

PL- condition :

- **Polyak-Lojasiewicz inequality:** suppose f satisfies:
  $|\nabla f(w)|^2 \geq c(f(w) - f(w^*))$ for all w

(true for SC functions, with $c = \frac{\mu}{8M}$:)

- "Global" condition, but can be satisfied for non-convex f

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

$$f(w_{t+1}) - f(w^*) \leq f(w_t) - f(w^*) - \frac{\eta}{2} \underbrace{\|\nabla f(w_t)\|^2}_{\geq c \cdot [f(w_t) - f(w^*)]}$$

$$\leq \underbrace{\left(1 - \frac{c\eta}{2}\right)\left(f(w_t) - f(w^*)\right)}.$$

PRECONDITIONING

- **Hessian** plays informal role in most GD analyses (M-smooth)

- "Directions" of Hessian can matter
- Optimal movement using second order information

IMPROVEMENTS, GENERALIZATIONS

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