THEORY OF MACHINE LEARNING

LECTURE 15

STRONG CONVEXITY, REGULARIZATION, STABILITY

Homework 2 is out. (zip. file submissions).
ok.

SUMMARY OF GRADIENT DESCENT

- Argmin f(x), over $x \in D$, where D is a convex domain
- Simple iterative algorithm ("first order")
- Fixed step-size

w_0

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

(project onto domain if needed).

low variance.

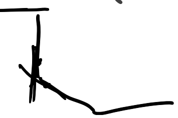
- Convergence with error $O(\frac{1}{\sqrt{T}})$ after T steps for any L-Lipschitz function

$T = 10^8$ $\bar{x}_t$ $f(\bar{x}_t) - f(x^*) \leq O\left(\frac{1}{\sqrt{T}}\right)$

- "Noisy gradient oracle" → stochastic gradient descent
- Error of O(1/T) for "smooth" convex functions (derivative is M-Lipschitz), assuming step size < $\frac{1}{2M}$

gradient.

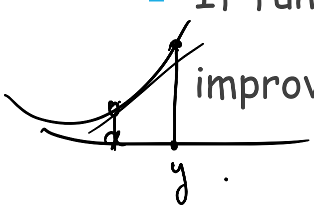
10^4

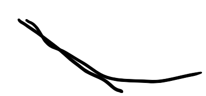


target error $= 10^{-4}$

- If function is also *strongly convex* with parameter μ, convergence bound improves to roughly $\exp(-\frac{\mu}{M}T)$ (extends to Polyak-Lojasiewicz)

10^{-4}





μ-strong convexity: $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \mu \|y - x\|^2$

Parallelism & Complexity.

$f : \mathbb{R}^d \to \mathbb{R}$.

d^2.

→ Optimization literature → gradients are "easy" but Hessians are "complex".

gradients → vector in $\mathbb{R}^d$

Hessians → $\mathbb{R}^{d \times d}$.

$\frac{\partial f}{\partial w_{t,i}}$

→ Parallelize grad. descent: → computing gradients can usually be parallelized.

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

→ Second-order methods: $w_{t+1} = \boxed{w_t - \eta \nabla f(w_t)}$.

$\boxed{H^{-1} \nabla f(w_t)}$. $\frac{1}{2}$.

Bengio-2015.

[Mahoney..].

[Training ~ imagenet.

$\log \log(\frac{1}{\varepsilon})$

$$f(x_{t+1}) - f(x^*) \leq \left(1 - \frac{1}{k}\right) \cdot \left(f(x_t) - f(x^*)\right)$$

$$f(x_{t+1}) - f(x^*) \leq \| f(x_t) - f(x^*) \|^2$$

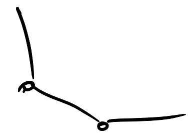
$\underset{nk.}{\overset{n \times n}{M}} \longrightarrow M^{-1}$ ($n \times n$)

$M^{-1} y$

Strong convexity based results are good only if M/μ is "small".

κ: condition number.

"OPTIMAL" BOUNDS



- Turns out: under just the Lipschitz assumption, $\frac{1}{\sqrt{T}}$ cannot be improved, at least with "sub-gradient oracle"

$1/T$.

10^{-2} iterations.

for err $= 10^{-4}$.

- **Smoothness:** purely assuming smoothness, can get rate of 1/T^2 (Nesterov 1983), this is optimal for all "gradient based" methods
- Can we use information *beyond* the gradient?

PRECONDITIONING

$$\nabla^2 f = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)$$

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

$\|\nabla^2 f\| \leq M.$

$x^T \nabla^2 f x \preceq x^T M I x$

$MI - \nabla^2 f \succeq 0.$

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

$$f(x_1, x_2, x_3) = x_1^2 + \frac{x_2^2}{4} + \frac{x_3^2}{9}$$

$(x_2-1)^2 \quad (x_3-1)^2$

$$\nabla f = \begin{pmatrix} 2x_1 \\ 2x_2/4 \\ 2x_3/9 \end{pmatrix} \rightarrow \text{Lipschitz const of } \nabla f \text{ is } 2.$$

$x^{(0)} = (1, 1, 1)$ should do gradient descent with step size $< \frac{1}{4}$.

$x_1^2 + (x_1 + x_2 - 1)^2 + \ldots$

$$\nabla f(x^{(0)}) = \frac{1}{4} \begin{pmatrix} 2 \\ 1/2 \\ 2/9 \end{pmatrix}$$

$$x^{(1)} = x^{(0)} - \frac{1}{4} \cdot \nabla f(x^{(0)}) = \left(\frac{1}{2} \right.$$

How much to move along diff. directions?

- what if we assume that function behaves like a quadratic in its neighborhood

$$f(x+\delta) \approx f(x) + \langle \delta, \nabla f(x)$$

$$\boxed{f(x+\delta) = f(x) + \langle \delta, \nabla f(x)\rangle + \frac{1}{2}\delta^T \left(\nabla^2 f(x)\right) \cdot \delta + \dots}$$

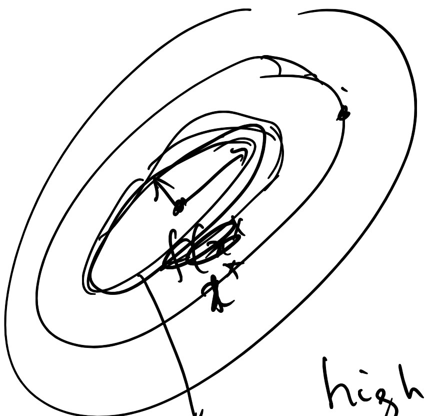
Hessian at x

$$f(x) + \delta f'(x) + \frac{\delta^2}{2!} f''(x)$$

$$ax^2 + bx + c$$

$$\frac{-b}{2a}$$

$$\delta = \frac{-f'(x)}{\cancel{2} f''(x)}$$



$\{x : f(x) = f(x^*) + 0.1\}$

high dim: $\delta = -\boxed{\left(\nabla^2 f(x)\right)^{-1} \nabla f(x)}$ (Newton's method)

IMPROVEMENTS, GENERALIZATIONS

for strongly convex f, instead of $e^{-\frac{T}{\kappa}}$

$e^{-T/\sqrt{\kappa}}$.

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

$\frac{1}{t}\left(\nabla f(x_1)\nabla f(x_1)^T + \dots + \dots\right)$

"first order approx" to second order methods.

STRONG CONVEXITY, MOTIVATION

- Saw that strong convexity leads to "faster" optimization

- Additional benefit – "stability" to small perturbation
  - Example of quadratic

$f(x) \to$ strongly convex $\longrightarrow$ x^2 $\quad$ 0

$g(x) = f(x) + \delta(x).$ $\longrightarrow$ $x^2 + \alpha x$ $\quad$ $-\frac{\alpha}{2}.$

$\arg\min_x f(x)$ is "close" to $\arg\min_x g(x).$

STABILITY OF LOSS MINIMIZATION

- Loss minimization with ‘n’ examples
- What happens if one example is “replaced”?

STABILITY IMPLIES GENERALIZATION

- Recall the notion of "generalization gap"
  - Can we phrase it in terms of stability?
- Stability versus "utility"!