UNIVERSITY OF CALIFORNIA, BERKELEY

Lecture 17 - EE127

Oscar Ortega

July 16, 2021

1 GRADIENT DESCENT

1.1 LEAST-SQUARES

Consider the least-squares problem:

$$\min_x f(x)$$

$$\text{Where } f(x) = \|Ax - b\|_2^2$$

$$= \langle Ax - b, Ax - b \rangle$$

$$= x^T A^T Ax - 2x^T A^T A + b^T b$$

$$\nabla_x = 2A^T Ax - 2A^T b$$

We know that by setting the gradient equal to 0 we can then solve for the optimal value of x which will be $x^* = (A^T A)^{-1} A^T b$, and solve for the minimization by plugging our minimizer into the objective function, but lets use gradient descent to solve the same problem:
Using our gradient descent update rule:

$$x^{k+1} = x^k - \eta \nabla f(x)$$

$$= (I - 2\eta A^T A)x^k - 2\eta A^T b$$

Here, one can see that the sequence will converge if $\|(I - 2\eta A^T A)\|_2 < 1$. There is a trade-off between computing the direct solution and arriving at the optimal solution via a gradient descent method.

- Solving for the optimum directly, $x^* = (A^T A)^{-1} A^T b$ can be computed in $O(n^3)$ time.
- Solving for the optimum via gradient descent can be computed in $O(n^2)k$, time where k is the number of steps needed until convergence.

2 Convergence

Definitions:

- $f : R^n \to R$ is an **L-Lipschits** function $L > 0$ if

$$\forall x, y : \|f(x) - f(y)\|_2 \leq L\|x - y\|_2$$

- $f : R^n \to R$ is a β **- smooth** function if f is continuous, differentiable, and

$$\forall x, y : \|\nabla f(x) - \nabla f(y)\|_2 \leq \beta\|x - y\|_2$$

Some things to know about the convergence of gradient descent iterations:

- f convex, L-Lipschits: convergence $\in O(\frac{1}{\sqrt{k}})$
- f convex, L-Lipschits, β - smooth: convergence $\in O(\frac{1}{\sqrt{k}})$
- f strongly-convex, L-Lipschits: convergence $\in O(\frac{1}{\sqrt{k}})$
- f strongly-convex, β - smooth: convergence $\in O(e^{-kc})$
- f non-convex, L-Lipschits, Armijo condition holds: stationarity to a local min $\in O(\frac{1}{\sqrt{k}})$

2.1 Backtracking-Algorithm

Start off with some really large step-size and scale back until we know convergence is achieved.

def Backtracking(β, α):

1. $s = s_{\text{init}}$
2. if $f(x_k + sv_k) > f(x) + \alpha\nabla f(x)^T v_k : s = \beta s$
3. else: go to line 2

Note that the condition on line 2 is just the Armijo condition, so what this algorithm seeks to find is the largest step-size that still satisfies the Armijo condition.
Theorem:
Let $f(x)$, be convex and L-lipschits, if T is the total number of steps taken and the learning rate is chosen as:

$$\gamma = \frac{\|x_1 - x^*\|_2}{L\sqrt{T}}$$

Then the following holds:

$$f(\frac{1}{T}\sum_{k=1}^{T} X_k) \leq \frac{\|x_1 - x^*\|L}{\sqrt{T}}$$

Lemma:
Given f : L-lipschits, continuous, and differentiable: $\|\nabla f\|_2^2 \leq L^2$
Proof:
By applying the Taylor expansion on f(x) at the point x_k, we have

$$f(x_k) - f(x^*) \leq \langle \nabla f(x_k), x_k - x^* \rangle$$

$$= \langle \frac{1}{\gamma}(x_k - x_{k+1}), x_k - x^* \rangle$$

$$= \frac{1}{2\gamma}(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2) + \gamma^2 \|\nabla f(x_k)\|_2^2$$

Using the lemma:

$$\leq \frac{1}{2\gamma}(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2) + \frac{\gamma}{2} L^2$$

Notice how if we sum the $(\|x_k - x^*\|_2^2 - \|x_{k+1} - x^*\|_2^2)$ terms, the sum is a telescoping sum.

By change of variable $\|x_1 - x^*\|_2^2$ to R:

$$\sum_{k=1}^{T} (f(x_k) - f(x^*) \leq \frac{R}{2\gamma} + \frac{L^2 T\gamma}{2} - \|x_k - x^*\|_2^2$$

Because norms are positive:

$$\sum_{k=1}^{T} (f(x_k) - f(x^*) \leq \frac{R}{2\gamma} + \frac{L^2 T\gamma}{2}$$

$$\frac{1}{T} \sum_{k=1}^{T} f(x_k) - f(x^*) \leq \frac{R}{2\gamma T} + \frac{\gamma L^2}{2}$$

Because f, is convex, for $\lambda \in [0,1]$ $f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$:

$$f(\frac{1}{T} \sum_{k=1}^{T} x_k) - f(x^*) \leq \frac{R}{2\gamma T} + \frac{\gamma L^2}{2}$$

Setting $\gamma = \frac{\|x_1 - x^*\|}{L\sqrt{T}}$

$$\leq \frac{\|x_1 - x^*\| L}{\sqrt{T}}$$