Lecture # 8 Gradient descent (unconstrained) E1 260

- Gradient descent for unconstrained problems

Lipschitz	$O\left(1/\varepsilon^2\right)$ $\varepsilon = 10^{-6}$ $T > 10^{12}$!!
Smooth	$O\left(1/\varepsilon\right)$ $T > 10^{6}$
Smooth & strongly convex	$O\left(\log\left(1/\varepsilon\right)\right)$ $T > 6$

$\eta = \frac{1}{L}$

- Quadratic function
  - Exact and backtracking line search

- Smooth convex functions : $O(1/\varepsilon)$: Sublinear convergence

$\rightarrow$ $f(\underline{y}) \leq f(\underline{x}) + \nabla f^T(\underline{x})(\underline{y}-\underline{x}) + \frac{L}{2}\|\underline{x}-\underline{y}\|^2$

$\forall\, \underline{x}, \underline{y} \in \mathrm{dom}\, f = X$

$\rightarrow$ A bound on optimality gap : $f(x) - f^*$

$f^* = f(\underline{x}^*)$ where x^* is a solution to min. $f(\underline{x})$

$$\frac{1}{2L}\|\nabla f(x)\|_2^2 \overset{(a)}{\leq} f(x) - f^* \overset{(b)}{\leq} \frac{L}{2}\|\underline{x} - \underline{x}^*\|_2^2$$

Gradient Descent : $\underline{x}_{t+1} = \underline{x}_t - \frac{1}{L}\nabla f(\underline{x}_t)$

with $\eta = \frac{1}{L}$

$$\Rightarrow \underline{x}_{t+1} - \underline{x}_t = -\frac{1}{L}\nabla f(\underline{x}_t)$$

$$f(\underline{x}_{t+1}) \leq f(\underline{x}_t) - \frac{1}{L}\|\nabla f(\underline{x}_t)\|_2^2 + \frac{1}{2L}\|\nabla f(\underline{x}_t)\|_2^2$$

(*) $$= f(\underline{x}_t) - \frac{1}{2L}\|\nabla f(\underline{x}_t)\|_2^2$$

$$\Rightarrow \frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla f(\underline{x}_t)\|_2^2 \leq \sum_{t=0}^{T-1} \left(f(\underline{x}_t) - f(\underline{x}_{t+1}) \right)$$

$$= f(\underline{x}_0) - f(\underline{x}_T)$$

(telescopic sum)

<u>Recall:</u>

$$\sum_{t=0}^{T-1} \left(f(\underline{x}_t) - f(x^*) \right) \leq \frac{\eta}{2} \sum_{t=0}^{T-1} \|\underline{g}_t\|_2^2 + \frac{1}{2\eta} \|\underline{x}_0 - \underline{x}^*\|^2$$

with $\eta = \frac{1}{L}$

$$\sum_{t=0}^{T-1} \left(f(\underline{x}_t) - f(\underline{x}^*) \right) \leq \frac{1}{2L} \sum_{t=0}^{T-1} \|\nabla f(\underline{x}_t)\|_2^2 + \frac{L}{2} \|\underline{x}_0 - x^*\|^2$$

$$\leq f(\underline{x}_0) - f(\underline{x}_T) + \frac{L}{2} \|\underline{x}_0 - \underline{x}^*\|^2$$

$$\Rightarrow \sum_{t=1}^{T} \left(f(\underline{x}_t) - f(\underline{x}^*) \right) \leq \frac{L}{2} \|\underline{x}_0 - \underline{x}^*\|_2^2$$

Since $f(\underline{x}_{t+1}) \leq f(\underline{x}_t) \quad \forall\, t \in [0, T]$

$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) = \left(\frac{1}{T}\sum_{t=1}^{T} f(\underline{x}_t)\right) - f(\underline{x}^*)$$

$$\geq f(\underline{x}_T) - f(x^*)$$

$$\Rightarrow f(\underline{x}_T) - f(\underline{x}^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(\underline{x}_t) - f(\underline{x}^*)$$

$$\leq \frac{L}{2T} \| \underline{x}_0 - \underline{x}^* \|_2^2 \quad ; T > 0$$

$$\frac{LR^2}{2T} = \varepsilon$$

with $R^2 = \| \underline{x}_0 - \underline{x}^* \|_2^2$

to obtain $\min\limits_{t=0 \ldots T-1} f(x_t) - f(x^*) \leq \varepsilon$

we need $T \geq \dfrac{R^2 L}{2\varepsilon}$

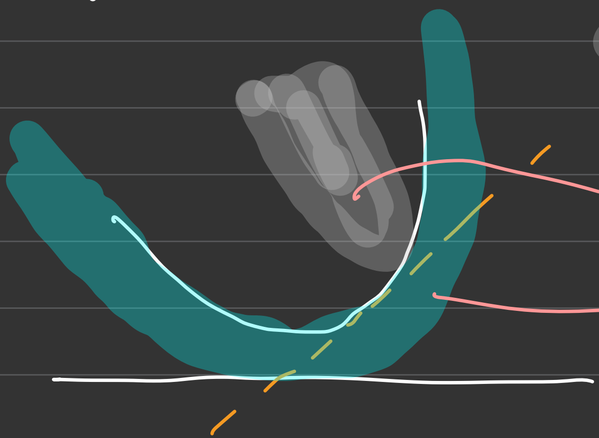
$$\boxed{\begin{array}{l}\text{Previously:} \\ T \geq \dfrac{R^2 B^2}{\varepsilon^2}\end{array}}$$

L - Smooth and μ - strongly convex functions:

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is μ - strongly convex and L - Smooth if

$$\frac{\mu}{2} \|\underline{x} - \underline{y}\|_2^2 \leq f(\underline{y}) - f(\underline{x}) - \nabla f^T(\underline{x})(\underline{y} - \underline{x}) \leq \frac{L}{2} \|\underline{x} - \underline{y}\|_2^2$$

Define $\kappa = \frac{L}{\mu}$ is the condition number



L - Smooth

μ - Strong convex

$\underline{x}^*$ is the minimizer

Gradient descent with a fixed step size

$$\underline{x}_{t+1} = \underline{x}_t - \frac{1}{L} \nabla f(\underline{x}_t)$$

Start with arbitrary $\underline{x}_0 \in \mathbb{R}^n$

Claim.

ⓐ Squared distances to $\underline{x}^*$ are geometrically decreasing

$$\|\underline{x}_{t+1} - \underline{x}^*\|^2 \leq \left(1 - \frac{\mu}{L}\right) \|\underline{x}_t - \underline{x}^*\|^2, \quad t \geq 0$$

$$\leq \left(1 - \frac{\mu}{L}\right)^t \|\underline{x}_0 - \underline{x}^*\|^2$$

ⓑ The error after T iterations is exponentially small in T:

$$f(\underline{x}_T) - f(\underline{x}^*) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\underline{x}_0 - \underline{x}^*\|^2; \quad T > 0$$

(a) Recall $\underline{g}_t = \nabla f(\underline{x}_t)$

$$\underline{g}_t^T(\underline{x}_t - \underline{x}^*) = \nabla f^T(\underline{x}_t)(\underline{x}_t - \underline{x}^*)$$

(from μ-strong convexity)
$$\geq f(\underline{x}_t) - f(\underline{x}^*) + \frac{\mu}{2}\|\underline{x}_t - \underline{x}^*\|_2^2$$

From vanilla analysis:

$$\underline{g}_t^T(\underline{x}_t - \underline{x}^*) = \frac{\eta}{2}\|\underline{g}_t\|^2 + \frac{1}{2\eta}\left[\|\underline{x}_t - \underline{x}^*\|^2 - \|\underline{x}_{t+1} - \underline{x}^*\|^2\right]$$

$$\Rightarrow f(\underline{x}_t) - f(\underline{x}^*) \leq \frac{1}{2\eta}\left[\eta^2\|\underline{g}_t\|^2 + \|\underline{x}_t - \underline{x}^*\|_2^2 - \|\underline{x}_{t+1} - \underline{x}^*\|_2^2\right] - \frac{\mu}{2}\|\underline{x}_t - \underline{x}^*\|_2^2$$

We have a bound on $\|\underline{x}_{t+1} - \underline{x}^*\|_2^2$:

$$\|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \leq \underbrace{2\eta\left[f(\underline{x}_t) - f(\underline{x}^*)\right] + \eta^2\|\underline{g}_t\|^2}_{\text{this disappear as shown next}} + (1-\mu\eta)\|\underline{x}_t - \underline{x}^*\|_2^2$$

For L-smooth convex function; for $\eta = \frac{1}{L}$:

$$f(\underline{x}^*) - f(\underline{x}_t) \leq f(\underline{x}_{t+1}) - f(\underline{x}_t) \leq -\frac{1}{2L} \|\nabla f(x_t)\|_2^2$$

$$2\eta \left[f(\underline{x}^*) - f(\underline{x}_t) \right] + \eta^2 \|\nabla f(\underline{x}_t)\|_2^2 \leq 0$$

$$\Rightarrow \|\underline{x}_{t+1} - \underline{x}^*\|_2^2 \leq (1 - \mu\eta) \|\underline{x}_t - \underline{x}^*\|_2^2$$

$$= \left(1 - \frac{\mu}{L}\right) \|\underline{x}_t - \underline{x}^*\|_2^2$$

$$\|\underline{x}_T - \underline{x}^*\|_2^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|\underline{x}_0 - \underline{x}^*\|_2^2$$

(b)

From smoothness:

$$f(x_T) - f(x^*) \le \underbrace{\nabla f^T(x^*)(x_T - x^*)}_{\nabla f(x^*) = 0} + \frac{L}{2}\|x_T - x^*\|_2^2$$

$$= \frac{L}{2}\|x_T - x^*\|_2^2$$

(b) $\Rightarrow$

$$\le \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \underbrace{\|x_0 - x^*\|_2^2}_{R^2} \quad \blacksquare$$

To find the number of iterations:

$$\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 = \varepsilon \Rightarrow \left(1 - \frac{\mu}{L}\right)^T = \frac{2\varepsilon}{R^2 L}$$

$$\Rightarrow \quad T \ln\left(1 - \frac{\mu}{L}\right) = \ln\left(\frac{2\varepsilon}{R^2 L}\right)$$

Since $\ln(1+x) \le x$

$$T\left(-\frac{\mu}{L}\right) \le \ln\left(\frac{2\varepsilon}{R^2 L}\right) \quad \Rightarrow \quad T \ge \frac{L}{\mu}\ln\left(\frac{R^2 L}{2\varepsilon}\right)$$

Summary :

Gradient descent with fixed step size

$$\eta = \frac{1}{L}$$

$\varepsilon = 10^{-6}$

Lipschitz	$O(1/\varepsilon^2)$	10^{12}
Smooth	$O(1/\varepsilon)$	10^{6}
Smooth & Strongly convex	$O(\log(1/\varepsilon))$	6

$\log(10^6)$

A similar result:

Suppose f is μ - strongly convex and L - smooth. Then gradient descent with $\eta_t = \eta = \dfrac{2}{\mu + L}$ satisfies

a. $$\|\underline{x}_T - \underline{x}^*\|_2^2 \le \left(\frac{\kappa - 1}{\kappa + 1}\right)^{2T} \|\underline{x}_0 - \underline{x}^*\|_2^2$$

b. $$f(\underline{x}_T) - f(\underline{x}^*) \le \frac{L}{2}\left(\frac{\kappa - 1}{\kappa + 1}\right)^{2T} \|\underline{x}_0 - \underline{x}^*\|_2^2$$

Homework 2.

Example: **Quadratic minimization:**

$$\underset{\underline{x}}{\text{minimize}} \quad f(\underline{x}) = \frac{1}{2} (\underline{x} - \underline{x}^*)^T Q (\underline{x} - \underline{x}^*)$$

$$Q \succ 0 \qquad \nabla f(\underline{x}) = Q(\underline{x} - \underline{x}^*)$$

$$\underline{x}_{t+1} - \underline{x}^* = \underline{x}_t - \underline{x}^* - \eta_t \nabla f(\underline{x}_t)$$
$$= (I - \eta_t Q)(\underline{x}_t - \underline{x}^*)$$

We have
$$\| \underline{x}_{t+1} - \underline{x}^* \|_2 \le \| I - \eta_t Q \| \, \| \underline{x}_t - \underline{x}^* \|$$

$$\| I - \eta_t Q \| = \max \{ |1 - \eta_t \lambda_1(Q)|, |1 - \eta_t \lambda_n(Q)| \}$$

η that yields $|1 - \eta_t \lambda_1(Q)| = |1 - \eta_t \lambda_n(Q)|$

$$\Rightarrow \quad \eta = \frac{2}{\lambda_1(Q) + \lambda_n(Q)}$$

So

$$\| I - \eta . Q \| = 1 - \frac{2\lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)} = \frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)}$$

$$\Rightarrow \quad \| \underline{x}_t - \underline{x}^* \|_2 \leq \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)} \right) \| \underline{x}_t - x^* \|_2$$

$$= \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) + \lambda_n(Q)} \right)^t \| \underline{x}_0 - \underline{x}^* \|_2$$

Exact line search:

$$\eta_t = \arg\min_{\eta \geq 0} f(\underline{x}_t - \eta \nabla f(\underline{x}_t))$$

$$\eta_t = \frac{\underline{g}_t^T \underline{g}_t}{\underline{g}_t^T Q \underline{g}_t}$$

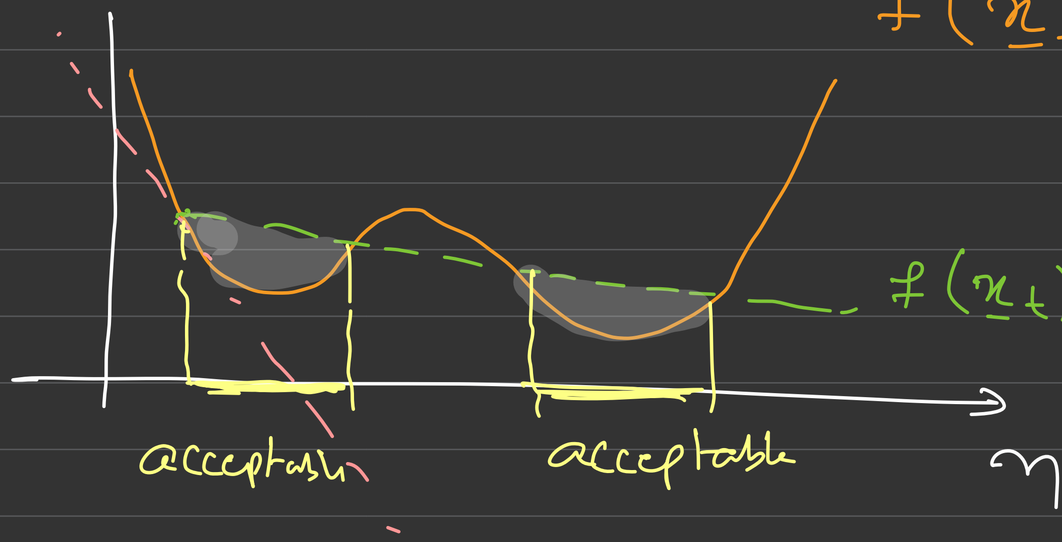
$$f(\underline{x}_t) - f(\underline{x}^*) \leq \left(\frac{\lambda_1(Q) - \lambda_n(Q)}{\lambda_1(Q) - \lambda_n(Q)} \right)^{2t} (f(x_0) - f(x^*))$$

(Homework 2)

- Convergence rate is not faster than fixed step size

Backtracking line search:

$$f(\underline{x}_t - \eta \nabla f(\underline{x}_t))$$



$$f(\underline{x}_t) - \alpha \eta \|\nabla f(\underline{x}_t)\|_2^2$$

$$f(\underline{x}_t) - \eta \|\nabla f(\underline{x}_t)\|^2$$

Armijo Condition: Ensures sufficient decrease in the objective value

$$0 < \alpha < 1$$

$$f(\underline{x}_t - \eta \nabla f(\underline{x}_t)) < f(\underline{x}_t) - \alpha \eta \|\nabla f(\underline{x}_t)\|_2^2$$

- $\eta = 1$, $0 < \alpha \leq 1/2$, $0 < \beta < 1$

while $f(\underline{x}_t - \eta \nabla f(\underline{x}_t)) > f(\underline{x}_t) - \alpha \eta \|\nabla f(\underline{x}_t)\|_2^2$

$\eta \leftarrow \beta \eta$

f is μ - strongly convex and L - smooth:

$$f(\underline{x}_t) - f(\underline{x}^*) \leq \left(1 - \min\left\{2\mu\alpha, \frac{2\beta\alpha\mu}{L}\right\}\right)^t \left(f(\underline{x}_0) - f(x^*)\right)$$