LECTURE 3

- PROOF NEWTON METHOD FOR CONVEX SETTINGS.

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

$$= x_t - (\nabla^2 f(x_t))^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_t - x^*))(x_t - x^*)\, d\tau - x^*$$

(BY TAYLOR'S THEOREM)

$$= (x_t - x^*) - (\nabla^2 f(x_t))^{-1} \int_0^1 \nabla^2 f(x^* + \tau(x_t - x^*))\, d\tau \cdot (x_t - x^*) - x^*$$

$$= (\nabla^2 f(x_t))^{-1} \cdot G_t \cdot (x_t - x^*)$$

WHERE $G_t = \int_0^1 \left(\nabla^2 f(x_t) - \nabla^2 f(x^* + \tau.(x_t - x^*))\right) d\tau$

OBSERVE THAT:

$$\| G_t \|_2 = \left\| \int_0^1 \left(\nabla^2 f(x_t) - \nabla^2 f(x^* + \tau(x_t - x^*))\right) d\tau \right\|_2$$

$$\leq \int_0^1 \| \nabla^2 f(x_t) - \nabla^2 f(x^* + \tau(x_t - x^*)) \|_2 \cdot d\tau$$

$$\leq \int_0^1 M \cdot \| x_t - x^* + \tau(x_t - x^*) \|_2 \, d\tau$$

$$= \frac{M \cdot \| x_t - x^* \|_2}{2}$$

MOREOVER, WE KNOW $\| \nabla^2 f(x) - \nabla^2 f(y) \|_2 \leq M \cdot \| x - y \|_2$

$$\Rightarrow \nabla^2 f(x) - M \cdot \|x - y\|_2 \cdot I \preccurlyeq \nabla^2 f(y) \preccurlyeq \nabla^2 f(x) + M \|x - y\|_2 \cdot I$$

THEN: $\nabla^2 f(x_t) \succcurlyeq \nabla^2 f(x^*) - M \cdot \|x_t - x^*\|_2 \cdot I \succcurlyeq (\mu - M \cdot \|x_t - x^*\|_2) \cdot I$

ASSUMING THAT $\|x_t - x^*\|_2 \leq \frac{\mu}{M} \longrightarrow \|\nabla^2 f(x_t)^{-1}\|_2 \leq (\mu - M\|x_t - x^*\|_2)^{-1}$

COMBINING ALL THE ABOVE:

$$\| x_{t+1} - x^* \|_2 \leq \| \nabla^2 f(x_t)^{-1} \|_2 \cdot \| G_t \|_2 \cdot \| x_t - x^* \|_2$$

$$\leq \frac{M \cdot \| x_t - x^* \|_2^2}{2 \cdot (\mu - M \cdot \|x_t - x^*\|_2)} \longrightarrow \text{THIS IS CALLED QUADRATIC CONVERGENCE.}$$

- PROOF OF HEAVY BALL METHOD.

REMEMBER THAT SIMPLE GRADIENT DESCENT, FOR SMOOTH AND STRONGLY CONVEX FUNCTIONS, IT SATISFIES:

$$\|x_t - x^*\|_2^2 \leq \left(1 - \frac{2}{K+1}\right)^t \|x_0 - x^*\|_2^2$$

FOR HEAVY BALL METHOD, WE OBSERVE:

$$\left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\|_2 = \left\| \begin{bmatrix} x_t - \eta \nabla f(x_t) + \beta(x_t - x_{t-1}) - x^* \\ x_t - x^* \end{bmatrix} \right\|_2$$

$$= \left\| \begin{bmatrix} x_t + \beta(x_t - x_{t-1}) - x^* \\ x_t - x^* \end{bmatrix} - \eta \begin{bmatrix} \nabla f(x_t) \\ 0 \end{bmatrix} \right\|_2$$

$$= \left\| \begin{bmatrix} (1+\beta)I & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} - \eta \begin{bmatrix} \nabla^2 f(z_t)(x_k - x^*) \\ 0 \end{bmatrix} \right\|$$

BLOCK MATRIX DIMENSIONS?

WHERE WE USED THE FACT: $\nabla f(x_t) = \nabla^2 f(z_k)(x_k - x^*)$

($\underline{\text{MEAN VALUE THEOREM}}$: LET $f: [\alpha, \beta] \longrightarrow \mathbb{R}$, DIFFERENTIABLE.
THEN, THERE EXISTS γ IN (α, β) SUCH THAT:

$$f'(\gamma) = \frac{f(\beta) - f(\alpha)}{\beta - \alpha}$$)

IN OUR CASE: $f'(\cdot) \longrightarrow \nabla^2 f(\cdot)$, $f(\cdot) \longrightarrow \nabla f(\cdot)$, AND WE KNOW $\nabla f(x^*) = 0$.
(ABUSE OF NOTATION)

$$\text{(CONT'D)} = \left\| \begin{bmatrix} (1+\beta)I - \eta \nabla^2 f(z_t) & -\beta I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} \right\|_2$$

$$\leq \left\| \begin{bmatrix} (1+\beta)I - \eta \nabla^2 f(z_t) & -\beta I \\ I & 0 \end{bmatrix} \right\|_2 \cdot \left\| \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} \right\|_2$$

LET US FOCUS ON THE FIRST TERM:

$$\left\| \begin{bmatrix} (1+\beta)I - \eta U \Lambda U^T & -\beta I \\ I & 0 \end{bmatrix} \right\|_2$$

$\nabla^2 f(z_k) \succ 0 \longrightarrow U \Lambda U^T$

$\underset{\text{EIGENVALUES.}}{\uparrow}$

$$\overset{\text{(INVARIANCE)}}{=} \left\| U^T \begin{bmatrix} (1+\beta)I - \eta U \Lambda U^T & -\beta I \\ I & 0 \end{bmatrix} U \right\|_2$$

$$= \left\| \begin{bmatrix} (1+\beta)U^T I U - \eta U^T U \Lambda U^T U & -\beta U^T I U \\ U^T I U & 0 \end{bmatrix} \right\|_2$$

$$= \left\| \begin{bmatrix} (1+\beta)I - \eta \Lambda & -\beta I \\ I & 0 \end{bmatrix} \right\|_2 = \max_i \left\| \begin{bmatrix} 1+\beta - \eta \lambda_i & -\beta \\ 1 & 0 \end{bmatrix} \right\|_2$$

THE EIGENVALUES OF 2x2 MATRICES ARE GIVEN BY:

$$p_i(\xi) = \xi^2 - (1 + \beta - \eta \lambda_i)\xi + \beta = 0.$$

"PLAYING" WITH THE VALUES OF β, ONE CAN CONCLUDE THAT:

$$\left\| \begin{bmatrix} (1+\beta)I - \eta \nabla^2 f(z_t) & -\beta I \\ I & 0 \end{bmatrix} \right\|_2 \leq \max \left\{ |1 - \sqrt{\eta \mu}|, |1 - \sqrt{\eta L}| \right\}$$

FOR $\beta := \max \left\{ |1 - \sqrt{\eta \mu}|, |1 - \sqrt{\eta L}| \right\}^2$

THUS; FOR $\eta = \dfrac{4}{(\sqrt{L} + \sqrt{\mu})^2}$:

$$\left\| \begin{bmatrix} x_{t+1} - x^* \\ x_t - x^* \end{bmatrix} \right\|_2 \leq \max \left\{ |1 - \sqrt{\eta \mu}|, |1 - \sqrt{\eta L}| \right\} \cdot \left\| \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} \right\|_2$$

$$= \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \cdot \left\| \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} \right\|_2 \leq \left(1 - \frac{1}{\sqrt{\kappa} + 1}\right) \cdot \left\| \begin{bmatrix} x_t - x^* \\ x_{t-1} - x^* \end{bmatrix} \right\|_2$$

COMPARE: $\left(1 - \dfrac{2}{\kappa + 1}\right)$ VS. $\left(1 - \dfrac{1}{\sqrt{\kappa} + 1}\right)$ ■

- WHAT IS THE DIFFERENCE BETWEEN $\frac{1}{t}$ AND $\frac{1}{t^2}$ IN IT. COMPLEXITY?

TO YIELD $f(x_t) - f(x^*) < \varepsilon$, WE NEED.

$$t > O\left(\frac{L \cdot \|x_0 - x^*\|_2^2}{\varepsilon}\right) \quad \text{FOR GRADIENT DESCENT}$$

$$t > O\left(\frac{L\|x_0 - x^*\|_2^2}{\sqrt{\varepsilon}}\right) \quad \text{FOR ACC. GRADIENT DESCENT.}$$

E.G. FOR $\varepsilon = 10^{-4} \longrightarrow 100\times$ FEWER STEPS!

- THEORY ON SGD

SETTING UP THE BACKGROUND:

1. SELECT $i_t \in [n]$ RANDOMLY (UNIFORM DIST.)
2. PERFORM $x_{t+1} = x_t - \eta \cdot \nabla f_{i_t}(x_t)$

OBSERVE THAT:

$$\mathbb{E}_{i_t}\left[\nabla f_{i_t}(x_t)\right] = \sum_{i=1}^{n} \mathbb{P}[i = i_t] \cdot \nabla f_i(x_t)$$

$$\overset{\text{UNIF.}}{=} \sum_{i=1}^{n} \frac{1}{n} \cdot \nabla f_i(x_t) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(x_t) = \nabla f(x_t)$$

I.E, $\nabla f_{i_t}(x_t)$ IS AN UNBIASED ESTIMATOR OF THE TRUE GRADIENT.

(DEFINITION OF "UNBIASEDNESS": THE DIFFERENCE BETWEEN EXPECTED VALUE AND TRUE VALUE)

STANDARD ASSUMPTIONS IN SGD: (BUT RESEARCHERS ARE WORKING ON REMOVING THEM)

IN ORDER TO LIMIT THE HARMFUL EFFECT OF STOCHASTICITY, WE REQUIRE THE VARIANCE OF $\nabla f_{i_t}(x_t)$ TO BE BOUNDED. I.E.

$$\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right] \leq M + M_f \cdot \|\nabla f(x_t)\|_2^2 \quad \text{FOR SOME } M, M_f \geq 0$$

(NOT OF IMPORTANCE FOR NOW TO DISCUSS ABOUT THESE)

OR $\leq C$ (IN SOME OTHER ANALYSES)

SGD FOR SMOOTH AND STRONGLY CONVEX f, WITH CONSTANT STEP SIZES

WE KNOW THAT:

$$f(x_{t+1}) \leq f(x_t) + \langle \nabla f(x_t), x_{t+1} - x_t \rangle + \frac{L}{2}\|x_{t+1} - x_t\|_2^2$$

$$= f(x_t) - \eta \langle \nabla f(x_t), \nabla f_{i_t}(x_t) \rangle + \frac{L\eta^2}{2}\|\nabla f_{i_t}(x_t)\|_2^2$$

TAKING EXPECTATION W.R.T i_t, FOR FIXED/GIVEN $i_1, \dots, i_{t-1}$

$$\mathbb{E}_{i_t}\left[f(x_{t+1})\right] \leq f(x_t) - \eta \langle \nabla f(x_t), \mathbb{E}_{i_t}[\nabla f_i(x_t)] \rangle + \frac{L\eta^2}{2} \cdot \mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right]$$

$$= f(x_t) - \eta \cdot \|\nabla f(x_t)\|_2^2 + \frac{L\eta^2}{2}\left(M + M_f \cdot \|\nabla f(x_t)\|_2^2\right) \Rightarrow$$

$$\mathbb{E}_{i_t}\left[f(x_{t+1}) - f(x^*)\right] = \left(f(x_t) - f(x^*)\right) - \left(\eta - \frac{L\eta^2 M_f}{2}\right)\|\nabla f(x_t)\|_2^2 + \frac{L\eta^2 M}{2} \quad (*)$$

BY STRONG CONVEXITY:

$$f(x_t) \leq f(x^*) + \langle \nabla f(x^*), x_t - x^* \rangle + \frac{1}{2\mu}\|\nabla f(x_t) - \nabla f(x^*)\|_2^2$$

$$\Rightarrow f(x_t) - f(x^*) \leq \frac{1}{2\mu} \cdot \|\nabla f(x_t)\|_2^2 .$$

THEN, FOR $\eta - \frac{L\eta^2 M_f}{2} \geq 0 \Rightarrow \eta \leq \frac{2}{L \cdot M_f}$, WE HAVE IN $(*)$:

$$\mathbb{E}_{i_t}\left[f(x_{t+1}) - f(x^*)\right] = \left(f(x_t) - f(x^*)\right) - 2\mu\left(\eta - \frac{L\eta^2 M_f}{2}\right) \cdot \left(f(x_t) - f(x^*)\right) + \frac{L\eta^2 M}{2}$$

$$= \left(1 - 2\mu\eta\left(1 - \frac{L\eta M_f}{2}\right)\right) \cdot \left(f(x_t) - f(x^*)\right) + \frac{L\eta^2 M}{2}$$

$$\left(\text{ASSUME } \eta = \frac{1}{LM_f}:\right) = (1 - \mu\eta)\left(f(x_t) - f(x^*)\right) + \frac{L\eta^2 M}{2}$$

REPEATING FOR i_t:

$$\mathbb{E}\left[f(x_{t+1}) - f(x^*)\right] \leq (1 - \mu\eta)^t \left(f(x_0) - f(x^*)\right) + \sum_{j=0}^{t} (1 - \mu\eta)^j \frac{L\eta^2 M}{2}$$

$$= (1 - \mu\eta)^t \left(f(x_0) - f(x^*)\right) + \frac{L\eta^2 M}{2} \cdot \frac{1 - (1 - \mu\eta)^{j+1}}{1 - 1 + \mu\eta}$$

$$\left(\text{ASSUMING } \eta \leq \frac{1}{\mu}:\right) \leq (1 - \mu\eta)^t \left(f(x_0) - f(x^*)\right) + \frac{L\eta M}{2\mu} \left(= O(\eta)\right)$$

THUS, FOR $\eta \leq \min\left\{\frac{1}{LM_f}, \frac{1}{\mu}\right\}$, WE GET THE ABOVE RESULT. ∎

SOME OBSERVATIONS:
1. FAST LINEAR CONVERGENCE WHEN FIRST PART ON RHS PREVAILS
2. AFTER THAT, CONVERGES AROUND A NEIGHBOURHOOD OF RADIUS $O(\eta)$

3. WHEN WE DO FULL GRADIENT DESCENT, $M = 0$, $M_f = 1$.
4. SMALLER STEP SIZES YIELD BETTER CONVERGING POINTS

(ANY COMMENTS / INTERPRETATIONS)?

SGD FOR SMOOTH AND STRONGLY CONVEX f, WITH DECREASING STEP SIZES

WE WILL CONSIDER A SIMPLER CASE FOR CLARITY PURPOSES.
ASSUME $\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right] \leq G^2$. WE CONSIDER:

$$x_{t+1} = x_t - \eta_t \cdot \nabla f_{i_t}(x_t), \quad \text{FOR } \eta_t = \frac{1}{\mu \cdot t}$$

WE KNOW THAT:

$$\mathbb{E}_{i_t}\left[\|x_{t+1} - x^*\|_2^2\right] = \|x_t - x^*\|_2^2 + \eta_t^2 \cdot \mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right] - 2\eta_k \left\langle \mathbb{E}_{i_t}\left[\nabla f_{i_t}(x_t)\right], x_t - x^* \right\rangle$$

BY STRONG CONVEXITY:

$$\langle \nabla f(x) - \nabla f(x^*), x - x^* \rangle = \langle \nabla f(x), x - x^* \rangle \geq \mu \cdot \|x - x^*\|_2^2$$

THEN,

$$\begin{aligned}
\mathbb{E}_{i_t}\left[\|x_{t+1} - x^*\|_2^2\right] &\leq \|x_t - x^*\|_2^2 + \eta_t^2 \cdot G^2 - 2\eta_t \langle \nabla f(x_t), x_t - x^* \rangle \\
&\leq \|x_t - x^*\|_2^2 + \eta_t^2 G^2 - 2\eta_t \mu \|x_t - x^*\|_2^2 \\
&= (1 - 2\eta_t \mu) \|x_t - x^*\|_2^2 + \eta_t^2 \cdot G^2
\end{aligned}$$

OBSERVE THAT: $\|x_1 - x^*\|_2^2 \leq \dfrac{\max\{\|x_1 - x^*\|_2^2, G^2/\mu^2\}}{1}$

WE WILL PROVE BY INDUCTION THAT:

$$\mathbb{E}\left[\|x_t - x^*\|_2^2\right] \leq \frac{\max\{\|x_1 - x^*\|_2^2, G^2/\mu^2\}}{t}$$

ASSUME IT HOLDS FOR t; THEN FOR $\Delta := \max\{\|x_1 - x^*\|_2^2, G^2/\mu^2\}$:

$$\begin{aligned}
\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] &\leq \left(1 - \frac{2}{t}\right) \cdot \mathbb{E}\left[\|x_t - x^*\|_2^2\right] + \frac{G^2}{\mu^2 \cdot t^2} \\
&\leq \left(1 - \frac{2}{t}\right) \cdot \frac{\Delta}{t} + \frac{G^2}{\mu^2 \cdot t^2}
\end{aligned}$$

$$\leq \left(\frac{1}{t} - \frac{2}{t^2}\right)\Delta + \frac{\Delta}{t^2} = \left(\frac{1}{t} - \frac{1}{t^2}\right)\Delta$$

$$\leq \frac{1}{t+1} \cdot \Delta. \qquad \blacksquare$$

- SGD FOR JUST SMOOTH CONVEX f, WITH DECREASING STEP SIZES

AGAIN, ASSUME $\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right] \leq G^2$. CHOOSE $\eta_t = \frac{c}{\sqrt{t}}$.

THEN, BY CONVEXITY WE HAVE:

$$\mathbb{E}\left[\langle \nabla f(x_t), x_t - x^* \rangle\right] \geq \mathbb{E}\left[f(x_t) - f(x^*)\right]$$

THEN, SIMILARLY TO THE ABOVE:

$$\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] \leq \mathbb{E}\left[\|x_t - x^*\|_2^2\right] - 2\eta_t \cdot \mathbb{E}\left[f(x_t) - f(x^*)\right] + \eta_t^2 G^2$$

$$\Rightarrow 2\eta_t \mathbb{E}\left[f(x_t) - f(x^*)\right] \leq \mathbb{E}\left[\|x_t - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] + \eta_t^2 \cdot G^2$$

SUMMING OVER $1, 2, \ldots, t$ ITERATIONS.

$$2\sum_{j=1}^{t} \eta_j \mathbb{E}\left[f(x_j) - f(x^*)\right] \leq \mathbb{E}\left[\|x_0 - x^*\|_2^2\right] - \mathbb{E}\left[\|x_{t+1} - x^*\|_2^2\right] + G^2 \cdot \sum_{j=1}^{t} \eta_j^2$$

$$\leq \mathbb{E}\left[\|x_0 - x^*\|_2^2\right] + G^2 \cdot \sum_{j=1}^{t} \eta_j^2 \Rightarrow$$

$$\sum_{j=1}^{t} \frac{\eta_j}{\sum_{\ell=1}^{t} \eta_\ell} \mathbb{E}\left[f(x_j) - f(x^*)\right] \leq \frac{\frac{1}{2}\mathbb{E}\left[\|x_0 - x^*\|_2^2\right] + G \cdot \sum_{j=1}^{t} \eta_j^2}{\sum_{j=1}^{t} \eta_j}$$

BY JENSEN INEQUALITY:

$$\sum_{j=1}^{t} \frac{\eta_j}{\sum_{\ell=1}^{t} \eta_\ell} \cdot f(x_j) \geq f\left(\sum_{j=1}^{t} \frac{\eta_j x_j}{\sum_{\ell=1}^{t} \eta_\ell}\right) := f(\tilde{x})$$

THUS:

$$\mathbb{E}\left[f(\tilde{x}) - f(x^*)\right] \leq \frac{\frac{1}{2}\mathbb{E}\left[\|x_0 - x^*\|_2^2\right] + G \cdot \sum_{j=1}^{t} \eta_j^2}{\sum_{j=1}^{t} \eta_j} \asymp O\left(\frac{\log t}{\sqrt{t}}\right)$$

FOR $\eta_t = \frac{c}{\sqrt{t}}$:

i. $\sum_{j=1}^{t} \eta_j^2 = \sum_{j=1}^{t} \frac{c^2}{j} \approx \log(t) + \frac{1}{2t}$

ii. $\sum_{j=1}^{t} \eta_j = \sum_{j=1}^{t} \frac{c}{\sqrt{j}} < 2\sqrt{t}$