LECTURE 2.

- ASSUME THE FOLLOWING UNCONSTRAINED OPTIMIZATION PROBLEM:

$$\min_{x \in \mathbb{R}^p} f(x)$$

ASSUMPTIONS: i) $f(x)$ IS DIFFERENTIABLE; I.E., $\nabla f(x)$ EXISTS, $\forall x$

ii) $f(x)$ IS SMOOTH; I.E. IT HAS LIPSCHITZ CONTINUOUS GRADIENTS:

$$\| \nabla f(x) - \nabla f(y) \|_2 \leq L \cdot \| x - y \|_2$$

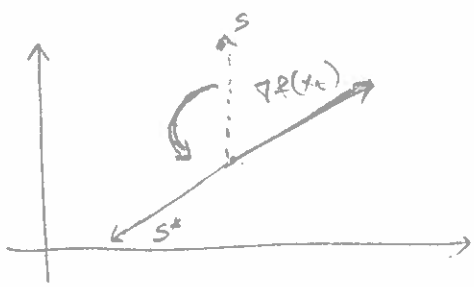
- HOW CAN WE SOLVE THIS PROBLEM? GRADIENT DESCENT.

CONSIDER THE 1ST-ORDER TAYLOR APPROXIMATION:

$$f(x_{t+1}) \approx f(x_t) + \langle \nabla f(x_t), \underbrace{x_{t+1} - x_t}_{s \in \mathbb{R}^p} \rangle .$$

IF WE WANT TO MINIMIZE THE RIGHT HAND SIDE, WE LOOK FOR THE DIRECTION THAT:

$$s^* \in \arg\min_{\|s\|_2 = 1} \langle \nabla f(x_t), s \rangle = - \frac{\nabla f(x_t)}{\| \nabla f(x_t) \|_2}$$



IN WORDS, THE DIRECTION WITH THE MAXIMAL DECREASE IN f IS THAT OF THE ANTIGRADIENT $-\nabla f(\cdot)$.

THUS:

INIT: CHOOSE $x_0 \in \mathbb{R}^p$

ITERATE: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t)$, $t = 0, 1, \ldots$

i) HOW WE CHOOSE STEP SIZE η_t?

ii) HOW WE CHOOSE x_0?

iii) HOW WE TERMINATE?

- FOR STEP SIZE, THERE ARE VARIOUS APPROACHES:
  i) $\eta_t = \eta$ (DEFINED BY USER)
  ii) $\eta_t = O\left(\frac{1}{t}\right)$ OR $O\left(\frac{1}{\sqrt{t}}\right)$ (DECREASING STEP SIZE)
  iii) $\eta_t = \arg\min_{\eta} f(x_t - \eta \nabla f(x_t))$ (OPTIMAL STEP SIZE).
  iv) OTHER SOPHISTICATED RULES: GOLDSTEIN-ARMIJO (OUT OF SCOPE).

- WHAT ABOUT INITIALIZATION? WE DON'T HAVE ENOUGH INFORMATION:
  $x_0 = 0$ OR $x_0 =$ RANDOM.

- WHAT ABOUT TERMINATION CRITERION?
  i) AFTER T (USER DEFINED) ITERATIONS.
  ii) WHEN $\|\nabla f(x_t)\|_2 \le \varepsilon$, FOR $\varepsilon > 0$ (USER-DEFINED)

- ANALYSIS: FOR $x_{t+1} = x_t - \eta_t \nabla f(x_t)$

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

(LIPSCHITZ PROPERTY)

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

$$= f(x_t) - \eta_t \|\nabla f(x_t)\|_2^2 + \frac{L}{2} \eta_t^2 \|\nabla f(x_t)\|_2^2$$

$$= f(x_t) - \eta_t \left(1 - \frac{\eta_t \cdot L}{2}\right) \cdot \|\nabla f(x_t)\|_2^2$$

OBSERVATION #1: IT SEEMS WE DECREASE THE OBJECTIVE BY SOME AMOUNT PER ITERATION. CAN WE DECREASE FOREVER UNTIL WE FIND x^*?

OBSERVATION #2: CAN WE FIND A GOOD STEP SIZE BY THIS EXPRESSION?

DEFINE: $g(\eta) = -\eta\left(1 - \frac{\eta L}{2}\right)$

$$g'(\eta) = 0 \Rightarrow \eta L - 1 = 0 \Rightarrow \boxed{\eta = \frac{1}{L}}$$

(RINGS A BELL?)

THEN: $f(x_{t+1}) \le f(x_t) - \frac{1}{2L} \|\nabla f(x_t)\|_2^2$

LET'S UNFOLD THIS RECURSION:

$$f(x_{T+1}) \leq f(x_T) - \frac{1}{2L} \|\nabla f(x_T)\|_2^2$$

$$f(x_T) \leq f(x_{T-1}) - \frac{1}{2L} \cdot \|\nabla f(x_{T-1})\|_2^2$$

$$\vdots$$

$$+ \quad f(x_1) \leq f(x_0) - \frac{1}{2L} \|\nabla f(x_0)\|_2^2$$

$$f(x^*) \leq f(x_{T+1}) \leq f(x_0) - \frac{1}{2L} \cdot \sum_{t=0}^{T} \|\nabla f(x_t)\|_2^2 \Rightarrow$$

$$\frac{1}{2L} \cdot \sum_{t=0}^{T} \|\nabla f(x_t)\|_2^2 \leq \underbrace{f(x_0) - f(x^*)}_{\text{CONSTANT / BOUNDED, DOES NOT DEPEND ON } T.}$$

IMPLIES THAT EVEN IF WE RUN FOR $T \to \infty$, THE ADDITION SHOULD BE SMALLER & SMALLER $\longrightarrow$ $\|\nabla f(x_t)\|_2 \to 0$.

HOWEVER, THIS DOES NOT IMPLY ANYTHING W.R.T. CONVERGENCE RATE.

$$\frac{1}{2L} \cdot (T+1) \cdot \min_t \|\nabla f(x_t)\|_2^2 \leq \frac{1}{2L} \sum_{t=0}^{T} \|\nabla f(x_t)\|_2^2 \leq f(x_0) - f(x^*)$$

$$\Rightarrow \min_t \|\nabla f(x_t)\|_2 \leq \sqrt{\frac{2L}{T+1}} \left(f(x_0) - f(x^*)\right)^{1/2} = O\left(\frac{1}{\sqrt{T}}\right)$$

HOW CAN WE USE THIS RESULT? FOR TERMINATION CRITERION

FIX $\varepsilon > 0$: FOR $\min_t \|\nabla f(x_t)\|_2 \leq \varepsilon$ WE REQUIRE:

$$T+1 \geq \frac{2L}{\varepsilon^2} \left(f(x_0) - f^*\right) \quad \text{ITERATIONS.}$$

$\left(\text{IN PRACTICE, } \frac{2L}{\varepsilon^2} \cdot f(x_0)\right)$.

- SOME BACKGROUND ON LOGISTIC REGRESSION.

"GIVEN A SAMPLE VECTOR $\alpha_i \in \mathbb{R}^p$ AND A BINARY CLASS $y_i \in \{\pm 1\}$ DEFINE THE CONDITIONAL PROBABILITY OF y_i GIVEN α_i AS:

$$\mathbb{P}\left[y_i \mid \alpha_i, x^* \right] \propto \frac{1}{1+\exp(-y_i \alpha_i^T x^*)} \longrightarrow \text{LOGISTIC FUNCTION."}$$

TAKING LOG-ML EXPRESSION, WE GET TO:

$$\min_x f(x) = \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \cdot \alpha_i^T x)\right)$$

GRADIENTS & HESSIAN OF $f(x)$:

$$\nabla f(x) = \frac{1}{n} \sum_{i=1}^{n} \nabla_x \left[\log\left(1+\exp(-y_i \cdot \alpha_i^T x)\right) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1+\exp(-y_i \cdot \alpha_i^T x)} \cdot \nabla_x \left[\exp(-y_i \alpha_i^T x) \right]$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{\exp(-y_i \alpha_i^T x)}{1+\exp(-y_i \alpha_i^T x)} \cdot -y_i \cdot \alpha_i^T$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{-y_i}{1+\exp(y_i \alpha_i^T x)} \cdot \alpha_i^T$$

$\nabla^2 f(x)$ (AS GRADIENT OF THE GRADIENT)

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{(1+\exp(y_i \alpha_i^T x))^2} \cdot \nabla_x \left[1+\exp(y_i \alpha_i^T x) \right] \cdot \alpha_i^T$$

$$= \frac{1}{n} \sum_{i=1}^{n} \underbrace{\frac{1}{(1+\exp(y_i \alpha_i^T x))^2} \exp(y_i \alpha_i^T x)}_{\text{SCALAR}} \cdot \underbrace{\alpha_i \cdot \alpha_i^T}_{\in \mathbb{R}^{n \times n}}$$

LIPSCHITZ GRADIENT CONTINUITY:

OBSERVE THAT: $\frac{1}{(1+\exp(\alpha))^2} \cdot \exp(\alpha) = \frac{1}{1+\exp(\alpha)} \cdot \frac{\exp(\alpha)}{1+\exp(\alpha)} = \frac{1}{1+\exp(\alpha)} \cdot \frac{1}{1+\exp(-\alpha)}$

DEFINE: $h(\alpha) = \frac{1}{1+\exp(-\alpha)} \in (0,1)$; ALSO OBSERVE THAT $h(-\alpha) = 1 - \alpha$.

THEN: $h(\alpha) \cdot h(-\alpha) \leq 0.25$.

GOING BACK TO HESSIAN DEFINITION:

$$\nabla^2 f(x) = \frac{1}{n} \sum_{i=1}^{n} h(y_i \alpha_i^T x) \cdot h(-y_i \alpha_i^T x) \cdot \alpha_i \alpha_i^T$$

$$\preceq \frac{1}{n} \sum_{i=1}^{n} 0.25 \cdot \alpha_i \alpha_i^T = \frac{1}{4n} \cdot A^T A$$

THUS: $\|\nabla^2 f(x)\|_2 \leq \frac{1}{4n} \cdot \max\{\text{eig}(A^T A)\} := L$ ∎

- DOES CONVEXITY HELP WITH GUARANTEES?

GRADIENT DESCENT: $x_{t+1} = x_t - \eta_t \cdot \nabla f(x_t) \longrightarrow$ WE WILL ASSUME $\eta_t = \eta$

ASSUMPTIONS: f HAS LIPSCHITZ CONTINUOUS GRADIENTS $\leftarrow$ (WE HAVE USED THIS BEFORE)

f IS CONVEX $\leftarrow$ (LET'S SEE WHAT THIS GIVES US)

WE HAVE: $\|x_{t+1} - x^*\|_2^2 = \|x_t - \eta \nabla f(x_t) - x^*\|_2^2$

$$= \|x_t - x^*\|_2^2 + \eta^2 \|\nabla f(x_t)\|_2^2 - 2\eta \langle \nabla f(x_t), x_t - x^* \rangle \quad (*)$$

EQUIVALENT FORMULATION OF GRADIENT LIPSCHITZ CONTINUITY FOR CONVEX FUNCTION

$$\frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_2^2 \leq \langle \nabla f(x) - \nabla f(y), x - y \rangle$$

USING THIS: $x = x^*$, $y = x_t$, $\nabla f(x^*) = 0$

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

$$\Rightarrow \langle \nabla f(x_t), x_t - x^* \rangle \geq \frac{1}{L} \|\nabla f(x_t)\|_2^2$$

$$\Rightarrow -2\eta \langle \nabla f(x_t), x_t - x^* \rangle \leq -\frac{2\eta}{L} \|\nabla f(x_t)\|_2^2$$

BACK TO $(*)$:

$$\|x_{t+1} - x^*\|_2^2 \leq \|x_t - x^*\|_2^2 - \eta\left(\frac{2}{L} - \eta\right) \cdot \|\nabla f(x_t)\|_2^2$$

$\leftarrow$ INTERPRETATION? WE DECREASE DISTANCE UNTIL WE FIND STATIONARY POINT.

THIS MEANS: $\|x_t - x^*\|_2^2 \leq \|x_0 - x^*\|_2^2$

WOULD WE HAVE SUCH A CONDITION IN NON-CONVEX SCENARIA?

WE ALSO KNOW THAT:

$$f(x_{t+1}) \le f(x_t) - \eta\left(1-\frac{L}{2}\eta\right)\|\nabla f(x_t)\|_2^2 \quad (**)$$

BY CONVEXITY:

$$f(x^*) \geq f(x_t) + \langle \nabla f(x_t), x^* - x_t \rangle$$

$$\Rightarrow \quad f(x_t) - f(x^*) \le \langle \nabla f(x_t), x_t - x^* \rangle \le \|x_t - x^*\|_2 \cdot \|\nabla f(x_t)\|_2$$
$$\le \|x_0 - x^*\|_2 \cdot \|\nabla f(x_t)\|_2$$

BACK TO $(**)$:

$$\left[f(x_{t+1}) - f(x^*)\right] \le \left[f(x_t) - f(x^*)\right] - \eta\left(1-\frac{L}{2}\eta\right)\|\nabla f(x_t)\|_2^2$$
$$\le \left[f(x_t) - f(x^*)\right] - \eta\left(1-\frac{L}{2}\eta\right) \cdot \frac{\left[f(x_t) - f(x^*)\right]^2}{\|x_0 - x^*\|_2^2}$$

DEFINE: $\Delta_t := f(x_t) - f(x^*)$

$$\Delta_{t+1} \le \Delta_t - \frac{\eta\left(1-\frac{L}{2}\eta\right)}{\|x_0-x^*\|_2^2} \cdot \Delta_t^2 = \Delta_t\left(1 - \frac{\eta\left(1-\frac{L}{2}\eta\right)}{\|x_0-x^*\|_2^2}\Delta_t\right) \Longrightarrow$$

$$\frac{\Delta_{t+1}}{\Delta_t} \le 1 - \frac{-\|-}{-\|-} \cdot \Delta_t \Rightarrow \frac{1}{\Delta_t} \le \frac{1 - \frac{-\|-}{-\|-}\Delta_t}{\Delta_{t+1}} \Longrightarrow$$

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{-\|-}{-\|-} \cdot \frac{\Delta_t}{\Delta_{t+1}} \geq \frac{1}{\Delta_t} + \frac{-\|-}{-\|-}$$

UNFOLDING THE RECURSION:

$$\frac{1}{\Delta_{t+1}} \geq \frac{1}{\Delta_0} + \frac{\eta\left(1-\frac{L}{2}\eta\right)}{\|x_0-x^*\|_2^2} \cdot (t+1)$$

SIMILARLY, OPTIMAL STEP SIZE $\eta = \frac{1}{L}$

THIS LEADS TO: $$f(x_t) - f(x^*) \le \frac{2L\left(f(x_0) - f^*\right) \cdot \|x_0 - x^*\|_2^2}{2L\|x_0-x^*\|_2^2 + t\left(f(x_0) - f^*\right)}$$

DEPENDS ON INITIAL CONDITIONS

WE CAN SIMPLIFY FURTHER: $f(x_0) \le f(x^*) + \frac{L}{2}\|x_0 - x^*\|_2^2 \longrightarrow$ WHICH CONDITION IS THIS?

$$\Rightarrow f(x_0) - f(x^*) \le \frac{L}{2}\|x_0 - x^*\|_2^2$$

USING IT IN OUR RESULT: $$f(x_t) - f(x^*) \le \frac{2L \cdot \|x_0 - x^*\|^2}{t+4} = O(1/t)$$

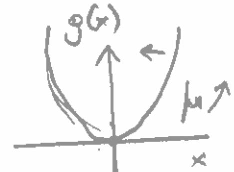
WHAT DOES THIS MEAN?

$$f(x_t) - f(x^*) \leq \varepsilon \xrightarrow{\text{REQUIRE}} \frac{2L \cdot \|x_0 - x^*\|_2^2}{t+4} \leq \varepsilon \Rightarrow t \geq \frac{2L \cdot \|x_0 - x^*\|_2^2}{\varepsilon}$$

$$\left(O\left(1/\varepsilon\right)\right)$$

- INTERPRETATION OF $\nabla^2 f(x) \succeq \mu \cdot I$

DEFINE: $g(x) = \frac{\mu}{2} \cdot \|x\|_2^2 \longrightarrow$



OR

$\nabla^2 g(x) = \mu \cdot I$. THUS $\nabla^2 f(x) \succeq \mu \cdot I \longrightarrow$ CURVATURE AT ANY POINT IS AT LEAST AS A $\frac{\mu}{2}\|x\|_2^2$ QUADRATIC.

- CONVERGENCE OF GRADIENT DESCENT ON STRONGLY CONVEX FUNCTIONS.

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

$$= \|x_t - x^*\|_2^2 + \eta^2 \cdot \|\nabla f(x_t)\|_2^2 - 2\eta \langle \nabla f(x_t), x_t - x^* \rangle \quad (*)$$

WE USE THE FACT:

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

THEN, $(*)$ BECOMES:

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

$$= \left(1 - \frac{2\eta\mu L}{\mu + L}\right)\|x_t - x^*\|_2^2 + \eta\underbrace{\left(\eta - \frac{2}{\mu + L}\right)}_{\leq 0 \text{ FOR } \eta \leq \frac{2}{\mu + L}}\|\nabla f(x_t)\|_2^2$$

$$\leq \left(1 - \frac{2\eta\mu L}{\mu + L}\right) \cdot \|x_t - x^*\|_2^2$$

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

$\eta = \frac{2}{\mu + L}$

$$\frac{2 \cdot \frac{2}{\mu + L} \cdot \mu L}{\mu + L} = \frac{4\mu L}{(\mu + L)^2} = \frac{4}{\frac{\mu}{L} + 2 + \frac{L}{\mu}} \leq \frac{2}{\mu + L}$$

FOR $\eta = \frac{2}{\mu + L}$:

$$1 - \frac{4\mu L}{(\mu + L)^2} \geq 0 \Rightarrow 4\mu L \leq (\mu + L)^2$$

$$\Rightarrow 4\mu L \leq \mu^2 + L^2 + 2\mu L$$

WHAT DOES THIS MEAN?

$$\|x_t - x^*\|_2^2 \leq \varepsilon \xrightarrow{\text{REQUIRES}} \left(1 - \frac{2\eta\mu L}{\mu+L}\right)^t \cdot \|x_0 - x^*\|_2^2 \leq \varepsilon.$$

$$\Rightarrow t \cdot \log\left(1 - \frac{2\eta\mu L}{\mu+L}\right) + \log\left(\|x_0 - x^*\|_2^2\right) \leq \log \varepsilon$$

$$\Rightarrow t \cdot \log\left(1 - \frac{2\eta\mu L}{\mu+L}\right) \leq \log \varepsilon / \|x_0 - x^*\|_2^2$$

$$\Rightarrow -t \cdot \log\left(1 - \frac{2\eta\mu L}{\mu+L}\right) \geq -\log \frac{\varepsilon}{\|x_0 - x^*\|_2^2}$$

$$\Rightarrow t \cdot \log \frac{1}{1 - \frac{2\eta\mu L}{\mu+L}} \geq \log \frac{\|x_0 - x^*\|_2^2}{\varepsilon}$$

$$\Rightarrow t \geq \frac{\log \|x_0 - x^*\|_2^2 / \varepsilon}{\log \frac{1}{1 - \frac{2\eta\mu L}{\mu+L}}} = O(\log 1/\varepsilon) \quad \blacksquare$$

- PL INEQUALITY.

WE KNOW THAT, BY LIPSCHITZ GRADIENT CONTINUITY:

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

PL DICTATES: $-\frac{1}{2}\|\nabla f(x_t)\|_2^2 \leq -\xi\left(f(x_t) - f(x^*)\right)$

THEN:

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

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

$$\Rightarrow f(x_{t+1}) - f(x^*) \leq \left(1 - \frac{\xi}{L}\right)^{t+1}\left(f(x_0) - f(x^*)\right) \quad \blacksquare$$

- CONVERGENCE OF PROJECTED GRADIENT DESCENT.

LET US ASSUME FOR SIMPLICITY THAT f IS L-SMOOTH AND μ-STRONGLY CONVEX.

WE KNOW THAT:

$$\|x_{t+1} - x^*\|_2^2 \leq \left(1 - \frac{2\eta\mu L}{\mu+L}\right)\|x_t - x^*\|_2^2$$

BY DEFINITION: $x_{t+1} = \Pi_C\left(x_t - \eta \nabla f(x_t)\right)$

THUS:

$$\| x_{t+1} - x^* \|_2^2 = \| \Pi_C\left(x_t - \eta \nabla f(x_t)\right) - x^* \|_2^2$$

$$= \| \Pi_C\left(x_t - \eta \nabla f(x_t)\right) - \Pi_C(x^*) \|_2^2$$

WHY IS THAT?

$$\leq \| x_t - \eta \nabla f(x_t) - x^* \|_2^2$$

$$\leq \dots \leq \left(1 - \frac{2\eta \mu L}{\mu + C}\right) \| x_t - x^* \|_2^2$$

(SAME CONVERGENCE RATES AS LONG AS $x^* \in \mathrm{INT}(C)$)
WE WILL NOT COVER THIS CASE.

SIMILAR EQUIVALENCE RESULTS HOLD FOR JUST SMOOTH CASES. ∎