LECTURE 9.

- PROOF OF SGD IN NON-CONVEX SCENARIA
  - ASSUMPTION #1: LIPSCHITZ GRADIENT CONTINUITY OF f:

$$\|\nabla f(x_1) - \nabla f(x_2)\|_2 \leq L \cdot \|x_1 - x_2\|_2$$

THIS FURTHER IMPLIES:

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

- ASSUMPTION #2: BOUNDED VARIANCE OF STOCHASTIC GRADIENTS:

$$\mathbb{E}_{i_t}\left[\|\nabla f_{i_t}(x_t)\|_2^2\right] \leq \sigma^2 \quad (*)$$

(EDIT: THIS PROOF ASSUMES STOCHASTIC GRADIENTS BOUNDED, SEE "STOCHASTIC VARIANCE REDUCTION FOR NONCVX OPT.")

- FROM RECURSION, $x_{t+1} = x_t - \eta \nabla f_{i_t}(x_t)$:

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

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

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

GIVEN x_t, AND TAKING EXPECTATION W.R.T. i_t, WE GET:

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

$$\Rightarrow$$

$$\eta \cdot \langle \nabla f(x_t), \mathbb{E}_{i_t}[\nabla f_{i_t}(x_t) \mid x_t] \rangle \leq \mathbb{E}_{i_t}[f(x_t) - f(x_{t+1}) \mid x_t] + \frac{\eta^2 L}{2} \cdot \sigma^2 \Rightarrow$$

$$\|\nabla f(x_t)\|_2^2 \leq \frac{\mathbb{E}_{i_t}[f(x_t) - f(x_{t+1}) \mid x_t]}{\eta} + \frac{\eta L \sigma^2}{2}$$

TAKING EXPECTATION W.R.T. x_t

$$\mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \leq \frac{\mathbb{E}[f(x_t) - f(x_{t+1})]}{\eta} + \frac{\eta L \sigma^2}{2}$$

UNFOLDING THE RECURSION OVER ALL ITERATIONS:

$$\mathbb{E}\left[\|\nabla f(x_1)\|_2^2\right] \leq \frac{\mathbb{E}\left[f(x_1)-f(x_2)\right]}{\eta} + \frac{\eta L \sigma^2}{2}$$

$$\mathbb{E}\left[\|\nabla f(x_2)\|_2^2\right] \leq \frac{\mathbb{E}\left[f(x_2)-f(x_3)\right]}{\eta} + \frac{\eta L \sigma^2}{2}$$

$$\cdots$$

$$+ \overline{\qquad\qquad\qquad\qquad\qquad\qquad}$$

$$\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \leq \frac{f(x_1) - \mathbb{E}\left[f(x_t)\right]}{\eta} + \frac{T\cdot \eta L \sigma^2}{2} \Rightarrow$$

$$T\cdot \min_{t=1..T} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \leq \frac{f(x_1) - \mathbb{E}\left[f(x_t)\right]}{\eta} + \frac{T \eta L \sigma^2}{2} \Rightarrow$$

$$\min_{t} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \leq \frac{f(x_1) - \mathbb{E}\left[f(x_t)\right]}{\eta \cdot T} + \frac{\eta L \sigma^2}{2}$$

ASSUME $f(x_1) - \mathbb{E}[f(x_t)] \leq D$. THEN, IF WE SET $\eta = \sqrt{\frac{D}{L\sigma^2/2} \cdot \frac{1}{T}}$

$$\min_{t} \mathbb{E}\left[\|\nabla f(x_t)\|_2^2\right] \leq \frac{D}{T} \cdot \frac{1}{\sqrt{\frac{D}{L\sigma^2/2}} \cdot T} + \sqrt{\frac{D}{L\sigma^2/2} \cdot \frac{1}{T}} \cdot \frac{L\sigma^2}{2}$$

$$= \frac{\sqrt{D\cdot L\sigma^2/2}}{\sqrt{T}} + \frac{\sqrt{D\cdot L\sigma^2/2}}{\sqrt{T}} = 2\sqrt{\frac{D L \sigma^2}{2\cdot T}}$$

IN WORDS: ASSUMING SMOOTHNESS, WE CAN APPROXIMATE A CRITICAL POINT IN $O\left(\frac{1}{\sqrt{T}}\right)$ ITERATIONS.

- DIAGONAL DERIVATION OF ADAGRAD & INTERPRETATION.

THE GENERAL FORM IS:

$$x_{t+1} = x_t - \frac{\eta}{\sqrt{\mathrm{diag}(B_t) + \varepsilon I}} \cdot \nabla f_{i_t}(x_t)$$

WHERE $B_t = \sum_{j=1}^{t} \nabla f_{i_j}(x_j) \nabla f_{i_j}(x_j)^T$

OBSERVE THAT:

$$\operatorname{diag}(B_t) = \begin{bmatrix} B_{t,(1,1)} & & & 0 \\ & B_{t,(2,2)} & & \\ & & \ddots & \\ 0 & & & B_{t,(p,p)} \end{bmatrix}$$

WHAT IS $B_{t,(q,q)}$? $\quad B_{t,(q,q)} = \sum_{j=1}^{t} \left(\nabla f_{i_j}(x_j)\right)_q^2 \longrightarrow$ SUM OF SQUARED GRADIENT WITH INDEX q.

THEN:

$$\frac{1}{\sqrt{\operatorname{diag}(B_t) + \varepsilon I}} = \begin{bmatrix} \frac{1}{\sqrt{B_{t,(1,1)} + \varepsilon}} & & & 0 \\ & \frac{1}{\sqrt{B_{t,(2,2)} + \varepsilon}} & & \\ & & \ddots & \\ 0 & & & \frac{1}{\sqrt{B_{t,(p,p)} + \varepsilon}} \end{bmatrix}$$

THUS:

$$x_{t+1,i} = x_{t,i} - \frac{\eta}{\sqrt{B_{t,(i,i)} + \varepsilon}} \cdot \left(\nabla f_{i_t}(x_t)\right)_i$$

INTERPETATION:

i) IF THE GRADIENT VALUES OF INDEX i ACROSS ITERATIONS IS LARGE

$\longrightarrow B_{t,(i,i)}$ IS LARGE $\longrightarrow \frac{1}{\sqrt{B_{t,(i,i)} + \varepsilon}}$ IS SMALL

ii) IF — // — — // — — // — — // — IS SMALL

$\longrightarrow B_{t,(i,i)}$ IS SMALL $\longrightarrow \frac{1}{\sqrt{B_{t,(i,i)} + \varepsilon}}$ IS LARGE

iii) INTUITION: TREAT EACH FEATURE MORE "DEMOCRATICALLY": IF A FEATURE APPEARS RARELY, WE USE A MORE AGGRESSIVE LEARNING RATE.

- EXPONENTIALLY WEIGHTED AVERAGES

$$V_t = \beta \cdot V_{t-1} + (1-\beta)\,\theta_t$$

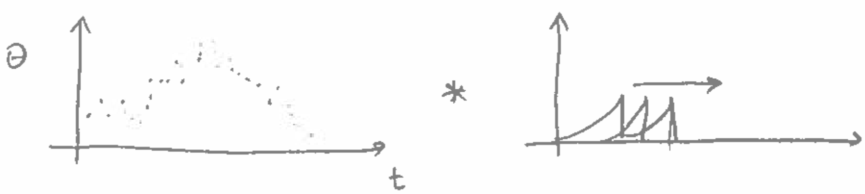
EXAMPLE:

$$V_{100} = \beta V_{99} + (1-\beta)\,\theta_{100}$$
$$V_{99} = \beta \cdot V_{98} + (1-\beta)\,\theta_{99}$$
$$\dots$$

CONSIDER: $\beta = 0.9$:

$$V_{100} = 0.1 \cdot \theta_{100} + 0.9 \cdot V_{99}$$
$$= 0.1 \cdot \theta_{100} + 0.9\left(0.1 \cdot \theta_{99} + 0.9\, V_{98}\right)$$
$$= 0.1 \cdot \theta_{100} + 0.9 \cdot 0.1 \cdot \theta_{99} + 0.9^2\left(0.1 \cdot \theta_{98} + 0.9 \cdot V_{97}\right)$$
$$= 0.1 \cdot \theta_{100} + 0.9 \cdot 0.1\, \theta_{99} + 0.9^2 \cdot 0.1 \cdot \theta_{98} + 0.9^3\, V_{97}$$
$$\dots$$

THIS IS EQUIVALENT TO:



IN WORDS: GIVE $(1-\beta)$ WEIGHT ON CURRENT TEMP;
GIVE $\beta(1-\beta)$ -11- ON PREVIOUS TEMP, ...

- BIAS CORRECTION

ASSUMING $\beta = 0.98$:

$$\boxed{V_0 = 0}$$

$V_1 = \cancel{V_0 \cdot 0.98}^{\,0} + 0.02\,\theta_1$ (THUS WE WEIGH WEIRDLY THE FIRST MEASUREMENTS)

$V_2 = 0.98\, V_1 + 0.02\,\theta_2 = 0.0192\,\theta_1 + 0.02\,\theta_1$ $\longrightarrow$ STILL WE DOWNGRADE A LOT THE ACTUAL TEMPS.

A WAY TO CORRECT THIS:

$\dfrac{V_t}{1-\beta^t}$: OBSERVATION: FOR t LARGE: $\dfrac{V_t}{1-\beta^t} \approx V_t$

FOR t SMALL: $(1-\beta)^t \overset{t=2}{=} 1 - (0.98)^t = 0.0396$

THEN: $\dfrac{V_2}{0.0396} = \dfrac{0.0196\,\theta_1 + 0.02\,\theta_1}{0.0396}$ (UPGRADES THE VALUE WE OBSERVE AND WEIGH TEMPS AT THE BEGINNING