LECTURE 8

- ANALYSIS OF HOGWILD!

TO MAKE THE ANALYSIS TRACTABLE, WE FOLLOW THE NEXT PROCEDURE:

i) EACH PROCESSOR SAMPLES e UNIFORMLY AT RANDOM

ii) — ll — COMPUTES GRADIENT OF f_e AT THE GIVEN POINT FROM PARAMETER SERVER.

iii) — ll — CHOOSES $v \in e$ UNIFORMLY AT RANDOM TO PERFORM THE UPDATE:

$$x \leftarrow x - \gamma \cdot |e| \cdot P_v^T \cdot G_e(x)$$

WHERE $P_v = b_v b_v^T$, AND b_v IS ALL-ZERO VECTOR EXCEPT AT POSITION v.

DISCLAIMER: THIS IS WASTEFUL COMPUTATIONALLY: WE KEEP ONLY ONE OF THE $|e|$ COMPUTED GRADIENT UPDATES. → THIS LEADS TO TRACTABLE ANALYSIS

ASSUMPTIONS: $f_e(\cdot)$ ARE CONVEX FUNCTIONS.

f HAS LIPSCHITZ CONTINUOUS GRADIENTS:

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

f IS STRONGLY CONVEX:

$$f(x') \geq f(x) + \langle \nabla f(x), x' - x \rangle + \frac{c}{2} \|x - x'\|_2^2$$

$\exists M > 0$ SUCH THAT:

$$\|G_e(x_e)\|_2 \leq M \quad \text{ALMOST SURELY } \forall x$$

(THIS IS RELATED TO f BEING A LIPSCHITZ FUNCTION)

ANALYSIS: BY STRONG CONVEXITY, WE HAVE:

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

RECALL THAT $k(j)$ IS THE STATE OF x_j WHEN WAS READ.

$$x_{j+1} = x_j - \gamma \cdot |e_j| \cdot P_{v_j} \cdot G_{e_j}(x_{k(j)})$$

SUBTRACTING x_* FROM BOTH SIDES, AND TAKING NORMS, WE GET:

$$\frac{1}{2}\| x_{j+1} - x^* \|_2^2 = \frac{1}{2}\| x_j - x^* \|_2^2 + \frac{1}{2}\gamma^2 \cdot |e_j|^2 \| P_{v_j} \cdot G_{e_j}(x_{k(j)}) \|^2$$
$$- \gamma \cdot |e_j| \langle P_{v_j} \cdot G_{e_j}(x_{k(j)}), x_j - x^* \rangle$$
$$= \frac{1}{2}\| x_j - x^* \|_2^2 + \frac{1}{2}\gamma^2 \cdot |e_j|^2 \| P_{v_j} \cdot G_{e_j}(x_{k(j)}) \|^2$$
$$- \gamma \cdot |e_j| \cdot \langle P_{v_j} \cdot G_{e_j}(x_{k(j)}), x_j - x_{k(j)} + x_{k(j)} - x^* \rangle$$
$$= \frac{1}{2}\| x_j - x^* \|_2^2 + \frac{1}{2}\gamma^2 \cdot |e_j|^2 \| P_{v_j} \cdot G_{e_j}(x_{k(j)}) \|^2$$
$$- \gamma \cdot |e_j| \cdot \langle P_{v_j} \cdot G_{e_j}(x_{k(j)}), x_j - x_{k(j)} \rangle$$
$$- \gamma |e_j| \cdot \langle P_{v_j} \cdot G_{e_j}(x_{k(j)}), x_{k(j)} - x^* \rangle$$
$$= \frac{1}{2}\| x_j - x^* \|_2^2 + \frac{1}{2}\gamma^2 \cdot |e_j|^2 \cdot \| P_{v_j} \cdot G_{e_j}(x_{k(j)}) \|^2$$
$$- \gamma \cdot |e_j| \cdot \langle P_{v_j} \cdot G_{e_j}(x_j), x_j - x_{k(j)} \rangle$$
$$- \gamma \cdot |e_j| \cdot \langle P_{v_j} \cdot (G_{e_j}(x_{k(j)}) - G_{e_j}(x_j)), x_j - x_{k(j)} \rangle$$
$$- \gamma \cdot |e_j| \cdot \langle P_{v_j} \cdot G_{e_j}(x_{k(j)}), x_{k(j)} - x^* \rangle$$

DEFINE $\alpha_j = \frac{1}{2}\mathbb{E}\left[\| x_j - x_* \|_2^2\right]$ (EXPECTATION TAKEN OVER RANDOM CHOISES)

THEN:

$$\alpha_{j+1} \leq \alpha_j - \gamma \cdot \mathbb{E}\left[\langle G_{e_j}(x_j), x_j - x_{k(j)} \rangle\right] \quad (**)$$
$$- \gamma \cdot \mathbb{E}\left[\langle G_{e_j}(x_{k(j)}) - G_{e_j}(x_j), x_j - x_{k(j)} \rangle\right] \quad (***)$$
$$- \gamma \cdot \mathbb{E}\left[\langle G_{e_j}(x_{k(j)}), x_{k(j)} - x^* \rangle\right] \quad (*)$$
$$+ \frac{1}{2}\gamma^2 \varrho^2 \cdot M^2$$

FOR THE TERM $(*)$, WE HAVE:

$$\mathbb{E}\left[\langle G_{e_j}(x_{k(j)}), x_{k(j)} - x^* \rangle\right] = \mathbb{E}\left[\mathbb{E}\langle G_{e_j}(x_{k(j)}), x_{k(j)} - x^* \rangle \mid e_1, e_2, \ldots, e_{j-1}, v_1, \ldots, v_{j-1}\right]$$
$$= \mathbb{E}\left[\langle \nabla f(x_{k(j)}), x_{k(j)} - x^* \rangle\right]$$

MOREOVER, BY STRONG CONVEXITY:

$$\mathbb{E}\left[\langle \nabla f(x_{k(j)}),\, x_{k(j)} - x^* \rangle\right] \geq c \cdot \alpha_{k(j)}$$

SIMILARLY FOR THE (**) TERM:

$$\mathbb{E}\left[\langle G_{e_j}(x_j),\, x_j - x_{k(j)} \rangle\right] = \dots = \mathbb{E}\left[\langle \nabla f(x_j),\, x_j - x_{k(j)} \rangle\right]$$

$$\geq \mathbb{E}\left[f(x_j) - f(k_{(j)})\right] + \frac{c}{2} \cdot \mathbb{E}\left[\| x_j - x_{k(j)} \|_2^2\right]$$

(FROM STRONG CONVEXITY)

THIS IS ONE PLACE WHERE ASYNCHRONY KICKS-IN:

$$\mathbb{E}\left[f(x_{k(j)}) - f(x_j)\right] = \sum_{i=k(j)}^{j-1} \mathbb{E}\left[f(x_i) - f(x_{i+1})\right] \longrightarrow$$

WE ASSUME ALL THESE x_i ARE COMPUTED STATES FOR x

REMEMBER, THIS INDICATES THE VALUE OF VARIABLE WHEN GRADIENT IS COMPUTED, AND MIGHT BE DIFFERENT THAN x_j

$$= \sum_{i=k(j)}^{j-1} \sum_{e \in E} \mathbb{E}\left[f_e(x_i) - f_e(x_{i+1})\right]$$

OBSERVE THAT: $f_e(x_i) - f_e(x_{i+1}) \leq \frac{1}{|E|} \langle G_e(x_i),\, x_i - x_{i+1} \rangle = \frac{\gamma}{|E|} \cdot \langle G_e(x_i), G_{e_i}(x_i) \rangle$

(CONVEXITY)

$$\leq \frac{\gamma}{|E|} \sum_{i=k(j)}^{j-1} \sum_{e \in E} \mathbb{E}\left[G_e(x_i)^T \underbrace{G_{e_i}(x_i)}_{\text{FIXED IN INNER LOOP}}\right]$$

$$\leq \frac{\gamma}{|E|} \cdot \sum_{i=k(j)}^{j-1} \sum_{e \in E} \rho M^2 = \frac{\gamma}{|E|} \cdot \sum_{i=k(j)}^{j-1} \rho \cdot |E| \cdot M^2$$

$$\leq \gamma \cdot \tau \cdot \rho \cdot M^2$$

THUS: $\mathbb{E}\left[\langle G_{e_j}(x_j),\, x_j - x_{k(j)} \rangle\right] \geq -\gamma\tau \cdot \rho \cdot M^2 + \frac{c}{2} \cdot \mathbb{E}\left[\| x_j - x_{k(j)} \|_2^2\right]$

FINALLY, FOR THE TERM (***) WE GET:

$$\mathbb{E}\left[\langle G_{e_j}(x_{k(j)}) - G_{e_j}(x_j),\, x_j - x_{k(j)} \rangle\right]$$

$$= \mathbb{E}\left[\sum_{i=k(j)}^{j-1} \langle G_{e_j}(x_{k(j)}) - G_{e_j}(x_j),\, x_{i+1} - x_i \rangle\right]$$

$$= \mathbb{E}\left[\sum_{i=k(j)}^{j-1} \langle G_{e_j}(x_{k(j)}) - G_{e_j}(x_j),\, \gamma |e_i| \cdot G_{e_i}(x_{k(i)}) \rangle\right]$$