Lecture #17 **Stochastic gradient methods** E1 260

- Empirical risk minimization

- Stochastic approximation of the gradient
  - unbiasedness
  - role of variance

- Lower bound $\underline{g}_t^T(\underline{x}_t - \underline{x}^*)$?

Wed. 20th 18:00 - 19:00 hrs (TA session)

Recall gradient descent :

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

$$\boxed{\underline{x}_{t+1} = \underline{x}_t - \eta_t \underline{g}_t} \quad ; \quad \underline{g}_t \in \partial f(\underline{x})$$

Motivation :

- $\underline{g}_t$ may be expensive to compute
- complete gradient $\underline{g}_t$ may not be available

So we use an stochastic or approximate version of $\underline{g}_t \in \partial f(\underline{x})$

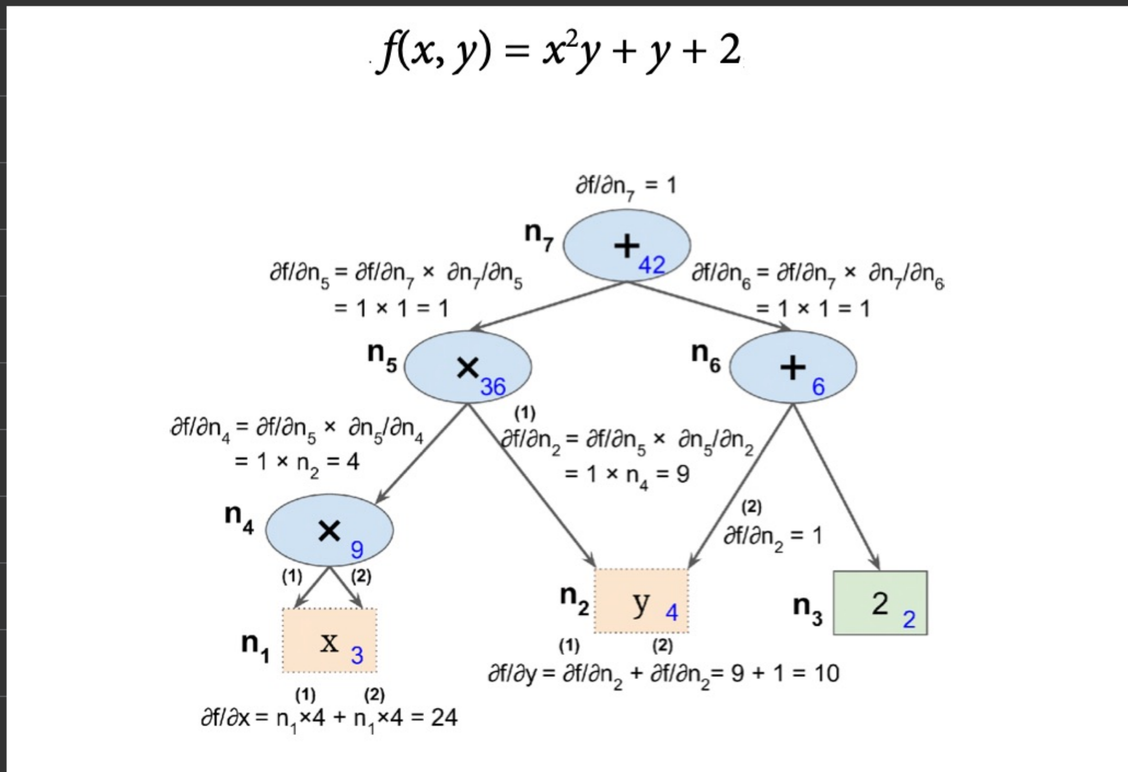
minimize $f(x)$

s. to $x \in C$

$\underline{x}$ → oracle (reverse mode auto diff) → $\nabla f(x)$

or

$\underline{g} \in \partial f(\underline{x})$



$$\underline{x} \longrightarrow \boxed{} \longrightarrow \underline{\tilde{g}}_x \quad \text{so that } E[\tilde{g}_x] \in \partial f(\underline{x})$$

(random variable)

Example:

① $$\underline{\tilde{g}}_x = \nabla f(x) + \underline{w} \quad ; \quad \underline{w} \text{ is zero-mean noise}$$

$$E[\tilde{g}_x] = E[\nabla f(x) + \underline{w}] = \nabla f(\underline{x})$$

② Random coordinate descent:

$$\nabla f(\underline{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{bmatrix}$$

$$x \longrightarrow \boxed{} \longrightarrow \tilde{g}_x = \begin{bmatrix} 0 \\ \vdots \\ \partial f / \partial x_j \\ \vdots \\ 0 \end{bmatrix} \cdot d \quad ; \quad \underline{x} \in \mathbb{R}^d$$

$$j \sim \text{unif}(1, \dots, d)$$

$$E_j[\tilde{g}_x] = \sum_{i=1}^{d} d \cdot \begin{bmatrix} 0 \\ \vdots \\ \partial f / \partial x_j \\ \vdots \\ 0 \end{bmatrix} \cdot \frac{1}{d} = \nabla f(\underline{x})$$

③ Stochastic programming:

$$\underset{\underline{x} \in C}{\text{minimize}} \quad F(\underline{x}) = \underbrace{\mathbb{E}\left[f(\underline{x} ; \xi) \right]}_{\substack{\text{Expected risk} \\ \text{population risk}}}$$

- ξ : randomness in the problem
- If $f(\underline{x} ; \xi)$ is convex for every ξ, then $F(\underline{x})$ is convex.

Empirical risk minimization:

Let $\{\underline{a}_i, y_i\}_{i=1}^{n}$ be n random data samples

ERM: $$\underset{\underline{x}}{\text{minimize}} \quad f(\underline{x}) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} f(\underline{x}; \{\underline{a}_i, \underline{y}_i\})}_{\text{Empirical risk}}$$

Regression problem: (more generally, any supervised learning)

$$f(\underline{x}; \{\underline{a}_i, y_i\}) = (\underline{a}_i^T \underline{x} - y_i)^2$$

prediction/hypothesis: $\underline{a}_i^T \underline{x}$

minimize expected loss: draw $j \sim \text{unif}(1, 2, \dots, n)$, then

$$E_j[f(\underline{x}; \{\underline{a}_j, y_j\})] = \sum_{j=1}^{n} f(\underline{x}; \{\underline{a}_j, y_j\}) \cdot \frac{1}{n}$$

$$\underset{\underline{x} \in C}{\text{minimize}} \quad f(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} f(x)$$

ERM: n is the number of data points

<u>Gradient</u> <u>descent</u>:

$$\underline{x}_{t+1} = \underline{x}_t - \eta \, \nabla \left(\frac{1}{n} \sum_{i=1}^{n} f_i(\underline{x}_t) \right)$$

$$= \underline{x}_t - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(\underline{x}_t)$$

When n is large, computing $\nabla f(x)$ is expensive (full pass over the data $\rightarrow$ memory issues)

$$\underline{x} \longrightarrow \boxed{\begin{matrix} j \sim \text{unif} \\ (1, \dots n) \\ \text{oracle} \end{matrix}} \longrightarrow \tilde{g}_x = \nabla f_j(\underline{x})$$

$$E_j \left[\nabla f_j(\underline{x}) \right] = \sum_{j=1}^{n} \nabla f_j(\underline{x}) \cdot \frac{1}{n} = \nabla f(\underline{x})$$

Stochastic gradient descent / Stochastic approximation:

$$\underline{x}_{t+1} = \underline{x}_t - \eta_t \, \tilde{g}(\underline{x}_t ; \xi_t)$$

where $\tilde{g}(\underline{x}_t ; \xi_t)$ is unbiased estimate of $\nabla f(\underline{x}_t)$, i.e.,

$$E\left[\tilde{g}(\underline{x}_t ; \xi_t)\right] = \nabla f(\underline{x}_t)$$

- Stochastic algorithm for finding a critical point $\underline{x}$ that obeys $\nabla F(\underline{x}) = 0$

  or for finding roots of $G(\underline{x}) = E\left[\tilde{g}(\underline{x}, \xi)\right]$

SGD for ERM

$$\underset{\underline{x}}{\text{minimize}} \quad f(\underline{x}) = \frac{1}{n} \sum_{i=1}^{n} f(\underline{x}; \{\underline{a}_i, y_i\})$$

for $t = 0, 1, \ldots$ do

Choose i_t uniformly at random

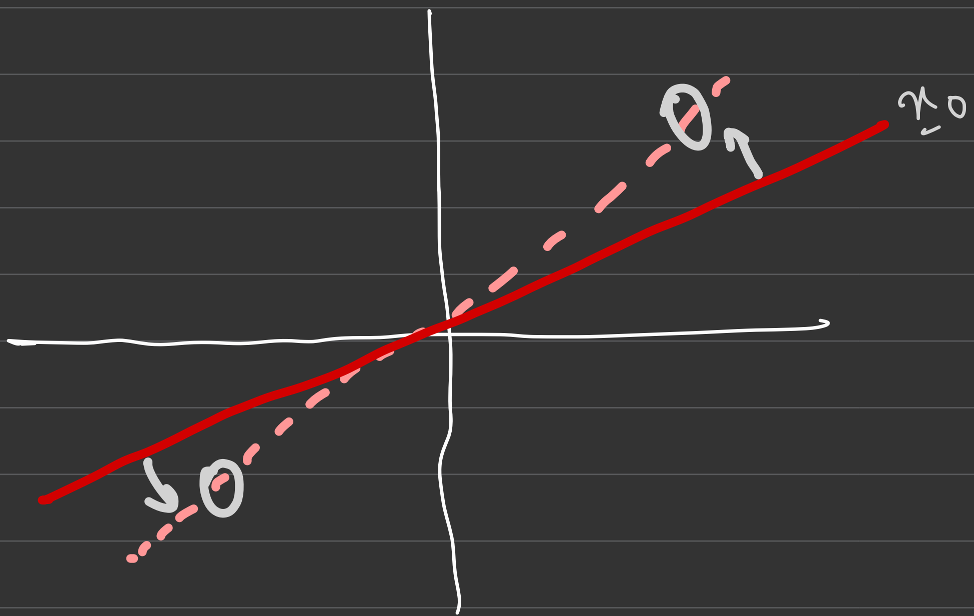
$$\underline{x}_{t+1} = \underline{x}_t - \eta_t \nabla_{\underline{x}} f_{i_t}(\underline{x}; \{\underline{a}_i, y_i\})$$

$y = \underline{a}_i^T \underline{x} + n$

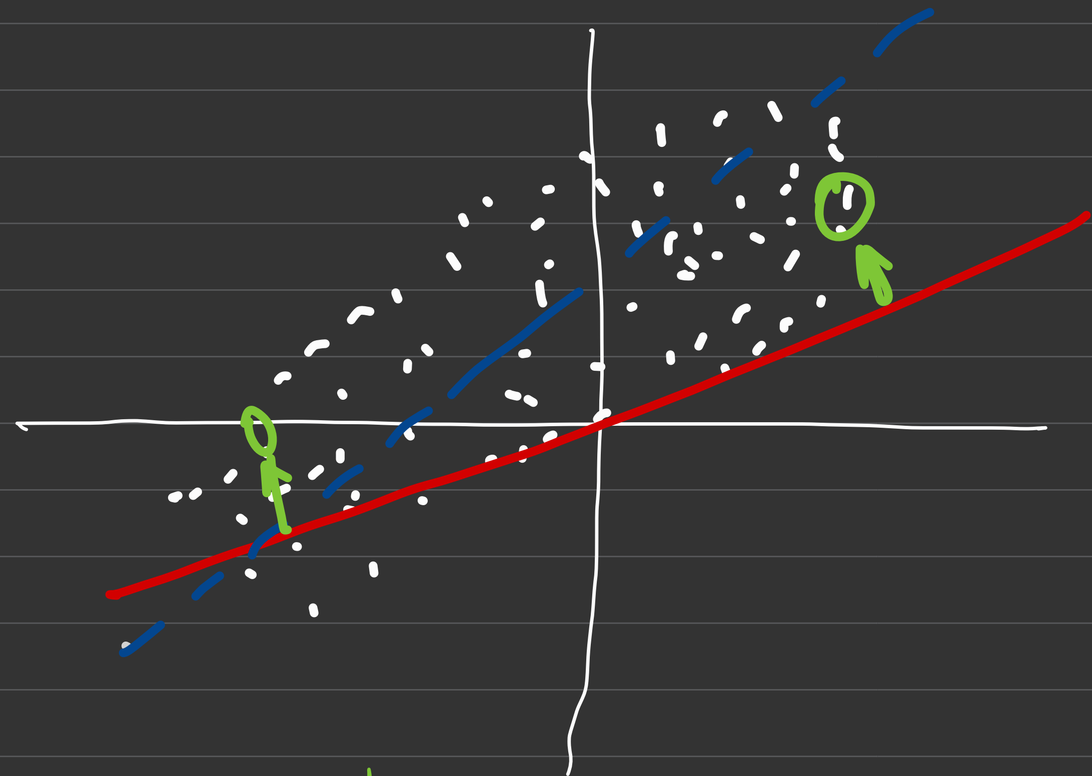
+ Exploits data more efficiently than batch methods

+ Fast initial improvement with low per-iteration cost (data usually has a lot of redundancy)

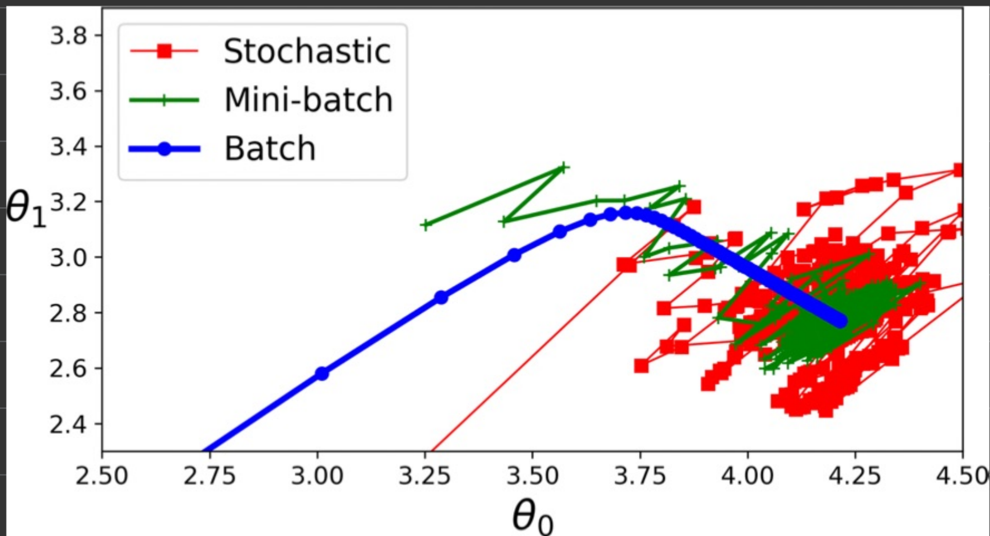
Role of variance:



low variance



high variance



Unbiasedness and the vanilla analysis:

Recall: In gradient descent, we could lower bound

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

but now we cannot as $\underline{\tilde{g}}_t$ may be far from being the true gradient.

- So inequality $f(\underline{x}_t) - f(\underline{x}^*) \leq \tilde{g}_t^T(\underline{x} - x^*)$

(from convexity) may not hold.

We have $\underbrace{\mathbb{E}[g_t \mid \underline{x}_t = \underline{x}]}_{} = \frac{1}{n}\sum_{j=1}^{n} \nabla f_j(\underline{x}) = \nabla f(\underline{x})$

Conditional expectation of g_t given the event $\{\underline{x} = x_t\}$.

$\forall\, \underline{x} \in \mathbb{R}^d$

$$\Rightarrow \mathbb{E}\left[g_t^T (\underline{x} - \underline{x}^*) \,\middle|\, \underline{x}_t = \underline{x} \right] =$$

$$\mathbb{E}\left[\underline{g}_t^T \,\middle|\, \underline{x}_t = \underline{x} \right] (\underline{x} - \underline{x}^*) = \nabla f^T(\underline{x}) (\underline{x} - \underline{x}^*)$$

- $\{ \underline{x}_t = \underline{x} \}$ can occur only for $\underline{x}$ in finite set X

$$\mathbb{E}\left[\underline{g}_t^T (\underline{x}_t - x^*) \right]$$

$$= \sum_{\underline{x} \in X} \mathbb{E}\left[g_t^T (\underline{x} - \underline{x}^*) \,\middle|\, \underline{x}_t = \underline{x} \right] \text{prob}(\underline{x}_t = \underline{x})$$

$$= \sum_{x \in X} \nabla f^T(x) (\underline{x} - \underline{x}^*) \text{ prob}(\underline{x}_t = \underline{x})$$

$$= \mathbb{E}\left[\nabla f^T(\underline{x}_t) (\underline{x}_t - \underline{x}^*) \right]$$

$$\Rightarrow \quad \mathbb{E}\left[\tilde{g}_t^T(\underline{x}_t - \underline{x}^*)\right] = \mathbb{E}\left[\nabla f^T(\underline{x}_t)(\underline{x}_t - \underline{x}^*)\right]$$

$$\geq \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*)\right]$$

So the lower bound holds in expectation.