Lecture # 18 **Stochastic gradient descent** E1 260

(contd.)

- Convergence analysis
- mini-batch variant

Stochastic gradient descent

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

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

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

ERM : minimize $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$

- Sample $i \in [n]$ uniformly at random
- $$\underline{x}_{t+1} = \underline{x}_t - \eta_t \underbrace{\nabla f_i(\underline{x}_t)}_{\tilde{g}_t \,=\, \text{Stochastic gradient}}$$

Bounded stochastic gradients:

- Same convergence rate as gradient descent method

Claim: Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex and differentiable function, $\underline{x}^*$ be a global minimum of f;

$\| \underline{x}_0 - \underline{x}^* \| \leq R$ and that $\mathbb{E}[\| \underline{g}_t \|^2] \leq B^2$ $\forall t$

Then stochastic gradient descent with constant step size $\eta = \dfrac{R}{B\sqrt{T}}$ yields

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[f(\underline{x}_t)] - f(\underline{x}^*) \leq \frac{RB}{\sqrt{T}}$$

Iteration complexity : $O\left(1/\varepsilon^2 \right)$ $\quad O\left(\frac{1}{\sqrt{T}}\right)$

Recall our vanilla analysis:

$$g_t^T(x_t - x^*) = \frac{\eta}{2}\|g_t\|^2 + \frac{1}{2\eta}\left(\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right)$$

Telescoping sum:

$$\sum_{t=0}^{T-1} g_t^T(x_t - x^*) = \frac{\eta}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\eta}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right)$$

$$\leq \frac{\eta}{2}\sum_{t=0}^{T-1}\|g_t\|^2 + \frac{1}{2\eta}\|x_0 - x^*\|^2$$

Taking expectation on both sides

$$\sum_{t=0}^{T-1}\mathbb{E}\left[\tilde{g}_t^T(x_t - x^*)\right] \leq \frac{\eta}{2}\sum_{t=0}^{T-1}\underbrace{\mathbb{E}\left[\|\tilde{g}_t\|^2\right]}_{\leq B^2} + \frac{1}{2\eta}\underbrace{\|x_0 - x^*\|^2}_{\leq R^2}$$

We have the lower bound:

$$\mathbb{E}\left[\tilde{g}_t^T(x_t - x^*)\right] \geq \mathbb{E}\left[f(x_t) - f(x^*)\right]$$

$$\sum_{t=0}^{T-1} \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*) \right] \leq \frac{\eta}{2} B^2 T + \frac{1}{2\eta} R^2$$

$$= q(\eta)$$

Choose η that minimize the upper bound:

$$\frac{1}{2} B^2 T - \frac{1}{2\eta^2} R^2 = 0$$

$$\eta = \frac{R}{B\sqrt{T}}$$

for which we have $O\left(1/\sqrt{T} \right)$ $\blacksquare$

$\Rightarrow$ This can be directly extended to "Projected Stochastic gradient descent"

- Sample $i \in [n]$
- $y_{t+1} = x_t - \eta_t \tilde{g}_t$
- $x_{t+1} = P_C(y_{t+1})$

Proj. SGD

$\min f(\underline{x})$

$\underline{x} \in C$

Strong Convexity:

- f is differentiable and μ strongly convex;

with a decreasing stepsize

$$\eta_t = \frac{2}{\mu(t+1)}$$

Stochastic gradient descent yields

$$\mathbb{E}\left[f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \underline{x}_t \right) - f(\underline{x}^*) \right] \leq \frac{2B^2}{\mu(T+1)}$$

$$B = \max_{t=1,\dots,T} \mathbb{E}\left[\|\underline{\tilde{g}}_t\| \right].$$

- We don't assume smoothness of f

  - diminishing step size

  (Similar to the analysis of subgradient)

Recall our vanilla analysis:

$$\underline{g}_t^T (\underline{x}_t - \underline{x}^*) = \frac{\eta}{2} \| g_t \|^2 + \frac{1}{2\eta} \left(\| \underline{x}_t - \underline{x}^* \|^2 - \| \underline{x}_{t+1} - \underline{x}^* \|^2 \right)$$

$$\mathbb{E}\left[\tilde{\underline{g}}_t^T (\underline{x}_t - \underline{x}^*) \right] = \frac{\eta_t}{2} \mathbb{E}\left[\| \tilde{\underline{g}}_t \|^2 \right] + \frac{1}{2\eta_t} \left[\mathbb{E}\left[\| \underline{x}_t - \underline{x}^* \|^2 \right] - \mathbb{E}\left[\| \underline{x}_{t+1} - \underline{x}^* \|^2 \right] \right]$$

Use strong convexity lower bound:

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

$$\geq \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*) \right] + \frac{\mu}{2} \mathbb{E}\left[\| \underline{x}_t - \underline{x}^* \|^2 \right]$$

$$\Rightarrow \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*)\right] \le \frac{B^2 \eta_t}{2} + \frac{1}{2}\left(\eta_t^{-1} - \mu\right)\mathbb{E}\left[\|\underline{x}_t - \underline{x}^*\|^2\right]$$

$$- \frac{\eta_t^{-1}}{2}\mathbb{E}\left[\|\underline{x}_{t+1} - \underline{x}^*\|^2\right]$$

Substituting $\eta_t = \dfrac{2}{\mu(t+1)}$:

$$t \cdot \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*)\right] \le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\Big[t(t-1)\mathbb{E}\left[\|\underline{x}_t - \underline{x}^*\|^2\right]$$

$$- (t+1)t\,\mathbb{E}\left[\|\underline{x}_{t+1} - \underline{x}^*\|^2\right]\Big]$$

$$\le \frac{B^2}{\mu} + \frac{\mu}{4}\Big[t(t-1)\mathbb{E}\left[\|\underline{x}_t - \underline{x}^*\|^2\right]$$

$$- (t+1)t\,\mathbb{E}\left[\|\underline{x}_{t+1} - \underline{x}^*\|^2\right]\Big]$$

Sum from $t = 1 \dots T$:

$$\sum_{t=1}^{T} t \cdot \mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*)\right] \leq \frac{B^2 T}{\mu} + \frac{\mu}{4}\left[0 - T(T+1)\,\mathbb{E}\left[\|\underline{x}_{t+1} - \underline{x}^*\|^2\right]\right]$$

$$\leq \frac{B^2 T}{\mu}$$

We have $\frac{2}{T(T+1)} \sum_{t=1}^{T} t = 1$.

$$\mathbb{E}\left[f\left(\frac{2}{T(T+1)} \sum_{t=1}^{T} t \cdot \underline{x}_t\right) - f(\underline{x}^*)\right] \leq \frac{2B^2}{\mu(T+1)}$$

$\Rightarrow$ ε - accuracy requires $O(1/\varepsilon)$ steps.

- Now, natural to ask if f is L-smooth and μ-strongly convex, will we get $O(\log(1/\varepsilon))$ (linear convergence) similar to the deterministic case.

Answer is <u>No</u>.

- Self-tuning property: $\nabla f(x) \to 0$ as $\underline{x} \to \underline{x}^*$

$\Rightarrow$ Allows a big step size $\left[\frac{1}{L} \text{ or } \frac{2}{\mu + L}\right]$

$\Rightarrow$ So far $\eta \sim \frac{1}{\sqrt{T}}$ or $\eta_t = \frac{2}{\mu(t+1)}$

- No self-tuning for SGD: $\mathbb{E}\left[\| \tilde{g}_x \|_2^2\right] \not\to 0$ as $\underline{x} \to \underline{x}^*$

  - SGD responds to every new sample
    - choose small steps close to the optimal

- μ - Strongly convex and L - Smooth

Suppose $\mathbb{E}\left[\| \tilde{g}_{\underline{x}} \|_2^2 \right] \leq \sigma_g^2 + C_g \| \nabla F(x) \|_2^2$

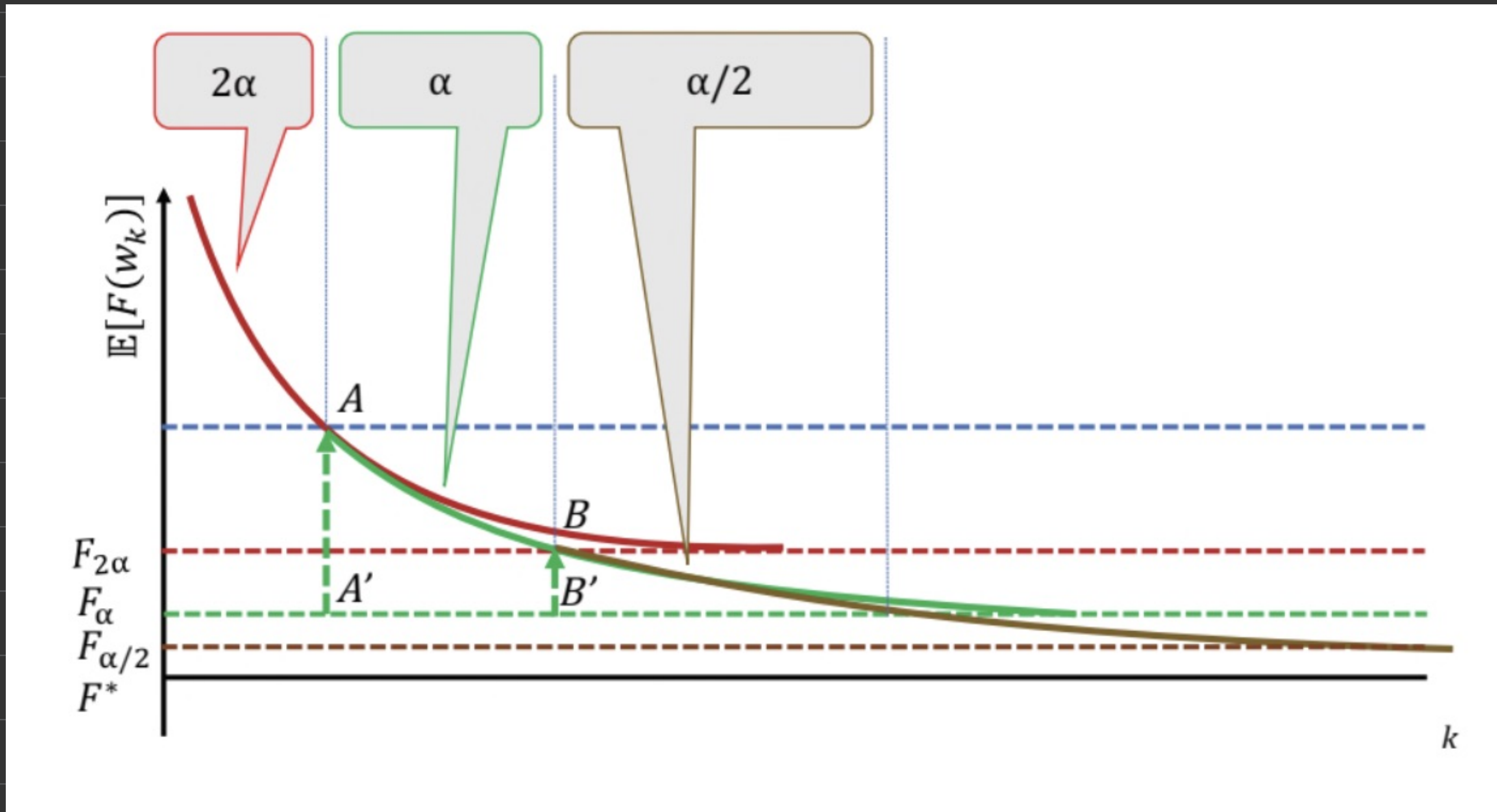
Then, SGD with fixed stepsize $\eta_t = \eta \leq \frac{1}{L C_g}$

yields

$$\mathbb{E}\left[f(\underline{x}_t) - f(\underline{x}^*) \right] \leq \frac{\eta L \sigma_g^2}{2\mu} + (1 - \eta\mu)^t \left[f(\underline{x}_0) - f(\underline{x}^*) \right]$$

- $\sigma_g = 0$: linear convergence
- converges to some neighborhood of $\underline{x}^*$

Practical trick:



When progress stalls, half the stepsize & repeat

Key question:

SGD with big stepsizes poorly suppresses noise. Larger stepsizes are needed for faster convergence.

$\boxed{\text{How to reduce the variance?}}$

Average iterates to reduce variance and improve convergence.

Mini-batch variants: (Tame the variance)

- Instead of choosing a single f_i from $\frac{1}{n}\sum_{i=1}^{n} f_i(\underline{x})$,

  let us pick several of them to form $\tilde{\underline{g}}_t$

- Let us pick B $\{f_i\}$: $f_1, f_2 \dots f_B$

  and average the gradients:

$$I_1, I_2 \dots I_j \sim \underset{B}{\text{uniform}}(1, \dots, n)$$

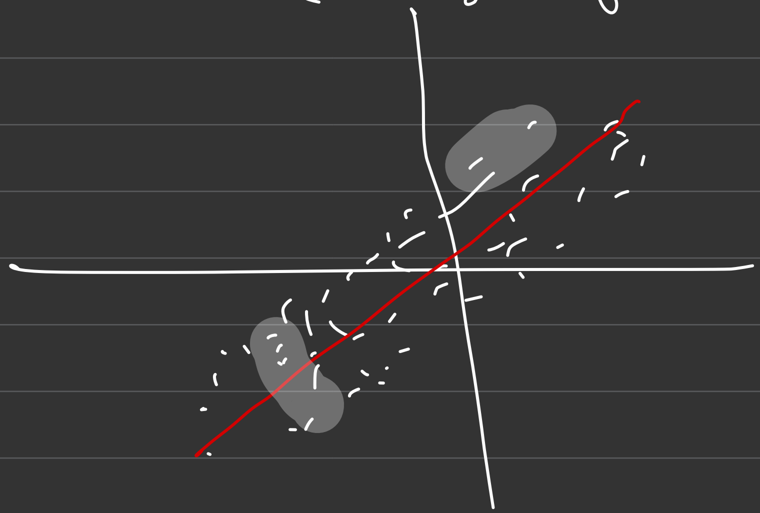
$$\underline{x}_{t+1} = \underline{x}_t - \eta \frac{1}{B} \sum_{j=1} \nabla f_{I_j}(\underline{x}_t)$$

- Stochastic gradient: $\mathbb{E}\left[\frac{1}{B}\sum_{j=1}^{B} \nabla f_{I_j}(\underline{x}_t)\right] = \frac{1}{B}\sum_{j=1}^{B} \mathbb{E}\left[\nabla f_{I_j}(\underline{x}_t)\right]$

$$= \frac{1}{B}\sum_{j=1}^{B} \nabla f(\underline{x}_t) = \nabla f(\underline{x}_t).$$

- $B = 1$, we have SGD $\leftarrow$
- $B = m$, we have full gradient descent

- Reduces variance: (average of independent r.v. reduces variance)



- <u>parallelization:</u>

$$\tilde{g}_t = \frac{1}{B} \sum_{j=1}^{B} \underbrace{\nabla f_{I_j}(\underline{x}_t)}$$

can be computed independently in parallel