Stochastic Gradient Descent

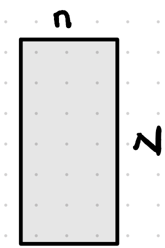
- motivation
- convergence in expectation

example : Large-scale least squares

Least-squares problem :

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad \frac{1}{2}\|Ax-b\|^2 = \frac{1}{2}\sum_{i=1}^{N} \underbrace{(a_i^T x - b_i)^2}_{\equiv f_i(x)}$$

$A =$



$$= \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_N^T \end{bmatrix}$$

gradient descent

$$x^+ = x - \alpha \nabla f(x) \qquad \nabla f(x) = A^T(Ax-b) = \sum_{i=1}^{N} \underbrace{a_i \cdot (a_i^T x - b_i)}_{\equiv \nabla f_i(x)}$$

may be prohibitive to form $\nabla f(x) = \sum_{i=1}^{N} \nabla f_i(x)$

- N is large
- data set $\{a_i, b_i\}_{i=1}^{N}$ is distributed

interpret objective as an expectation

objective may be interpreted as an expectation over samples $i = 1 \cdots N$ that occur with equal probability $1/N$:

$$\text{minimize} \quad f(x) := \frac{1}{N} \sum_{i=1}^{N} f_i(x) = \mathbb{E}_i f_i(x)$$

by linearity of the gradient operator and finiteness of sum:

$$\nabla f(x) = \frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x) = \mathbb{E}_i \nabla f_i(x)$$

randomly sample a small batch of observations $B \subseteq \{1, \cdots, N\}$. Then

$$g_B(x) := \frac{1}{|B|} \sum_{i \in B} \nabla f_i(x) \qquad \text{and} \qquad \mathbb{E}_B g_B(x) = \nabla f(x)$$

is a stochastic approximation to $\nabla f(x)$

stochastic gradient descent (SGD)

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \quad f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x)$$

gradient descent $\qquad x_{k+1} = x_k - \alpha_k \nabla f(x_k)$

stochastic gradient descent $\qquad x_{k+1} = x_k - \alpha_k g_k$

where $\quad g_k := \frac{1}{|B_k|} \sum_{i \in B_k} \nabla f_i(x_k) \qquad B_k := \left\{ \begin{array}{l} \text{batch of uniformly random} \\ \text{iid samples from } \{1 \dots N\} \end{array} \right\}$

- step length α_k often called "Learning rate" in this context.

- need to assume mean-squared error in stochastic approx is bounded:

$$\mathbb{E}\left[\|g_k - \nabla f(x_k)\|^2\right] = \mathbb{E}\left[\|g_k\|^2\right] - \|\nabla f(x)\|^2 \leq \sigma^2$$

for some $\sigma > 0$ fixed $\forall k$. (larger sample size $\Rightarrow$ smaller σ)

convergence in expectation (simplified version w/ constant step length)

by descent lemma: $f_k := f(x_k)$ and $\nabla f_k := \nabla f(x_k)$

$$f_{k+1} \le f_k + \nabla f_k^T (x_{k+1} - x_k) + \frac{L}{2} \| x_{k+1} - x_k \|^2$$

sgd step: $x_{k+1} = x_k - \alpha g_k \implies x_{k+1} - x_k = -\alpha g_k$

$$f_{k+1} \le f_k - \alpha \nabla f_k^T g_k + \frac{L}{2} \| -\alpha g_k \|^2$$

take expectations over both sides:

$$\begin{aligned}
\mathbb{E} f_{k+1} &\le \mathbb{E} \left[f_k - \alpha \nabla f_k^T g_k + \frac{\alpha^2 L}{2} \| g_k \|^2 \right] \\
&\le \mathbb{E} f_k - \alpha \mathbb{E} \| \nabla f_k \|^2 + \frac{\alpha^2 L}{2} \left(\sigma^2 + \mathbb{E} \| \nabla f_k \|^2 \right) \\
&= \mathbb{E} f_k - \alpha \left(1 - \frac{\alpha L}{2} \right) \mathbb{E} \| \nabla f_k \|^2 + \frac{\alpha^2 \sigma^2 L}{2} \\
&\le \mathbb{E} f_k - \frac{\alpha}{2} \mathbb{E} \| \nabla f_k \|^2 + \frac{\alpha^2 \sigma^2 L}{2}
\end{aligned}$$

(last line holds when steplength $\alpha < 1/L$)

from previous slide:

$$\mathbb{E} f_{k+1} \leq \mathbb{E} f_k - \frac{\alpha}{2} \mathbb{E} \|\nabla f_k\|^2 + \frac{\alpha^2 \sigma^2 L}{2}$$

Suming over $k = 0, 1, 2, \dots, T$ and recursing:

$$\mathbb{E} f_K \leq f(x_0) - \frac{\alpha}{2} \sum_{k=0}^{T-1} \mathbb{E} \|\nabla f_k\|^2 + \frac{\alpha^2 \sigma^2 L T}{2}$$

rearrange and divide both sides by $\alpha T/2 > 0$

$$\frac{1}{T} \sum_{k=0}^{T} \mathbb{E} \|\nabla f(x_k)\|^2 \leq \frac{2\,(f(x_0) - f^*)}{\alpha T} + \underbrace{\frac{\alpha \sigma L}{2}}_{\text{error term}}$$

compare to previous deterministic analysis