OPTIMIZATION FOR MACHINE LEARNING · STOCHASTIC GRADIENT

October 12, 2023

Program : Today : Basics of SG
Oct 19 : Theory of SG methods
Oct 23 : Advanced SG methods

INTRODUCTION TO STOCHASTIC GRADIENT

→ So far : $\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x)$ $\quad f \; C^1 \Rightarrow$ can query a gradient ∇f at every point

Gradient descent

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \qquad \alpha_k > 0$$

→ In ML, f generally has a specific structure that depends on some data

$\mathcal{A}$: input space $\qquad \mathcal{Y}$: output space

<u>Goal</u>: Compute a mapping between $\mathcal{A}$ and $\mathcal{Y}$ such that for a given data distribution $\mathcal{D}$ on $(\mathcal{A}, \mathcal{Y})$, the mapping maps a to y for any $(a,y) \sim \mathcal{D}$

⇒ Actual task: find the <u>best</u> <u>possible</u> mapping

<u>Quantify</u> <u>how</u> <u>good</u> <u>a</u> <u>mapping</u> <u>is</u> <u>at</u> <u>fitting</u> <u>the</u> <u>data</u>

- Expected risk : For a mapping $h : \mathcal{A} \to \mathcal{Y}$,

$$R(h) = \underset{\text{probability}}{\mathbb{P}}_{(a,y)\sim\mathcal{D}}\left[h(a) \neq y \right] = \underset{\text{expected value}}{\mathbb{E}}\left[\mathbb{1}(h(a) \neq y) \right]$$

$$\mathbb{1}(E) = \begin{cases} 1 & \text{if } E \text{ happens} \\ 0 & \text{otherwise} \end{cases}$$

⇒ A natural optimization problem

$$\underset{h \in \mathcal{H}}{\text{minimize}} \; R(h)$$

where $\mathcal{H}$ is some function space

↳ In general, one cannot design algorithms to solve this problem in its generic.

↳ <u>Modifications</u>

- $\mathcal{H}$ can be hard to optimize over $\Rightarrow$ parameterize

$$h \in \mathcal{H} \Rightarrow h \in \{ h(\cdot\,; x) \mid x \in \mathbb{R}^d \}$$

- $\mathbb{E}_{(a,y)\sim D}[\cdot]$ not known in practice $\Rightarrow$ approximate with samples

$$\mathbb{E}_{(a,y)\sim D}[\cdot] \Rightarrow \frac{1}{m} \sum_{i=1}^{m} (\underset{\uparrow \atop (a_i, y_i) \sim D}{\cdot})$$

- $\mathbb{1}(h(a) \neq y)$ "0-1 loss" is not easy to optimize

$\Rightarrow$ replace with a continuous loss (typically convex)

$$\mathbb{1}(h(a) \neq y) \Rightarrow \ell(h(a), y)$$
$$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$

Ex) $\ell(h,y) = \|h-y\|^2$
$\ell(h,y) = \|h-y\|_1$

With these changes, the problem becomes

$$\underset{x \in \mathbb{R}^d}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^{m} \underbrace{\ell(h(a_i; x), y_i)}_{f_i(x)}$$

Variables $\rightarrow$ $x \in \mathbb{R}^d$

Average over data samples

$f_i(x)$: loss at the data point (a_i, y_i)

"Empirical Risk Minimization"

We set $$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x) = \frac{1}{m} \sum_{i=1}^{m} \ell(h(a_i; x), y_i)$$

↳ If every f_i is C^1, then f is C^1 and

$$\forall x \in \mathbb{R}^d, \quad \nabla f(x) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(x)$$

and so gradient descent (GD) performs the iteration

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} \nabla f_i(x_k)$$

⚠ In a "big data" setting, computing $\nabla f(x_k)$ is expensive because it involves computing all $\nabla f_i(x_k)$ for $i = 1..n$, i.e. going through the entire dataset.

n samples $\{(a_1, y_1), \dots, (a_n, y_n)\}$

In addition, it may be possible to compute a better point than x_k without looking at all data points.

(1) STOCHASTIC GRADIENT ALGORITHM

↳ Aka stochastic gradient descent (SGD)

Setup: $\underset{x \in \mathbb{R}^d}{\text{minimize}} \; f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x)$ $\qquad f_i : \mathbb{R}^d \to \mathbb{R}$ C^1

"Finite-sum problem"

Key assumption: Every value of f_i or its gradient ∇f_i is computed using the ith sample of a dataset with n elements

Stochastic Gradient iteration (Robbins & Monro, 1954)

$\forall k \geq 0$, $$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) \qquad \alpha_k > 0$$

↳ "stochastic gradient"

where i_k is drawn randomly in $\{1, \dots, m\}$

↳ One iteration of SG (Stochastic Gradient) accesses 1 data point only

$\Rightarrow$ m times cheaper than one iteration of GD (which accesses m data points)

More general: Batch stochastic gradient

Iteration: $$x_{k+1} = x_k - \alpha_k \left(\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k) \right) \qquad \alpha_k > 0$$

cardinality of S_k

where S_k is a set of indices drawn randomly (with or without replacement) in $\{1, \dots, m\}$

$\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k)$: "Batch stochastic gradient"

- $|S_k| = 1 \Rightarrow$ SG
- $|S_k| = m$ and draws without replacement $\Rightarrow$ GD aka batch GD

$$\frac{1}{|S_k|} \sum_{i \in S_k} \nabla f_i(x_k) = \frac{1}{m} \sum_{i=1}^{m} \nabla f_i(x_k) = \nabla f(x_k)$$

For fixed batch size $|S_k| = m_b$ $\forall k$, we distinguish two regimes outside of $m_b = 1$ and $m_b = m$:

- $m_b \gg 1$: large batch regime, per-iteration cost is (much) larger than SG

  much larger than

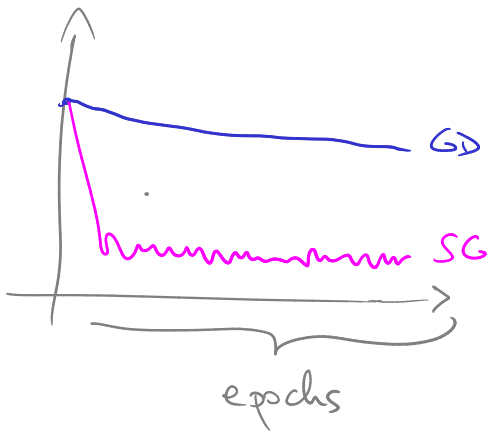
  $\Rightarrow$ m_b accesses to data points per iteration

  $\rightarrow$ In practice, can be similar to GD in terms of speed of convergence

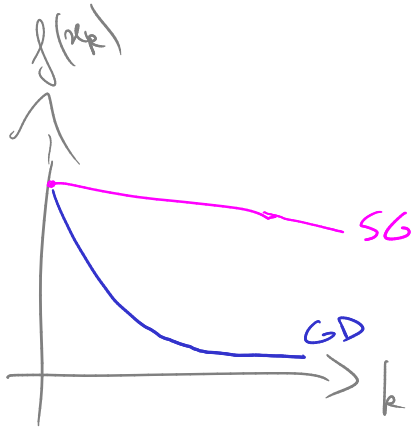
- $1 < m_b \ll n$: mini-batch regime
  - per-iteration cost larger than SG but much smaller than GD
  - practical performance can be better than that of SG

(3) Comparing SG, GD and batch SG

Typical plot



Not the plot you want



↳ Iteration count is not the right metric to compare algorithms for solving our problem of interest

↳ A better metric: Number of epochs

Def. An epoch is the cost of n accesses to a point in a dataset of size n.

1 iteration of GD $= 1\ \nabla f(\cdot) = \nabla f_1, \dots, \nabla f_n = n$ accesses to data points

1 iteration of GD costs 1 epoch

1 iteration of SG costs $\frac{1}{n}$ epoch

1 iteration of batch SG with fixed batch size m_b costs $\frac{m_b}{n}$ epoch(s)

4 Towards theory for SG

We consider again

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

<u>Assumptions</u>:

- f_i is C^1 $\forall i \in \{1, \dots, n\}$
- f is $C^{1,1}_L$, i.e. C^1 with L-Lipschitz continuous gradient

∇f is L-Lipschitz continuous with $L > 0$

$\iff \forall (x, w) \in (\mathbb{R}^d)^2, \quad \|\nabla f(x) - \nabla f(w)\| \leq L \|x - w\|$

<u>Proposition</u>: Under these assumptions, for any $(x, w) \in (\mathbb{R}^d)^2$,

$$f(w) \leq f(x) + \nabla f(x)^T (w - x) + \frac{L}{2} \|w - x\|^2 \qquad (1)$$

↳ For GD, apply (1) with $\forall x = x_k$ and $w = x_k - \alpha_k \nabla f(x_k)$,

$$f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - \alpha_k \|\nabla f(x_k)\|^2 + \frac{L}{2} \alpha_k^2 \|\nabla f(x_k)\|^2$$

- Allows to identify good values for α_k that guarantee $f(x_k - \alpha_k \nabla f(x_k)) < f(x_k)$
- One of the main ingredients to show asymptotic

or non-asymptotic convergence properties of GD

↳ For SG, we apply (1) with $x = x_k$ and $w = x_k - \alpha_k \nabla f_{i_k}(x_k)$

$$f(x_k - \alpha_k \nabla f_{i_k}(x_k)) \leq f(x_k) + \nabla f(x_k)^T (x_k - \alpha_k \nabla f_{i_k}(x_k) - x_k) + \frac{L}{2} \| x_k - \alpha_k \nabla f_{i_k}(x_k) - x_k \|^2$$

$$= f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L}{2} \| -\alpha_k \nabla f_{i_k}(x_k) \|^2$$

$$= f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L \alpha_k^2}{2} \| \nabla f_{i_k}(x_k) \|^2$$

$$(2) \quad \underbrace{f(x_k - \alpha_k \nabla f_{i_k}(x_k)) \leq f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L \alpha_k^2}{2} \| \nabla f_{i_k}(x_k) \|^2}_{\text{inequality between random variables that depend (in particular) on } i_k}$$

<u>Assumption</u>: The random indices $\{i_k\}_k$ are independent

Under that assumption, we take the expectation with respect to i_k on both sides of (2)

$$\mathbb{E}_{i_k}\left[f(x_k - \alpha_k \nabla f_{i_k}(x_k)) \right] \leq \mathbb{E}_{i_k}\left[f(x_k) - \alpha_k \nabla f(x_k)^T \nabla f_{i_k}(x_k) + \frac{L}{2} \alpha_k^2 \| \nabla f_{i_k}(x_k) \|^2 \right]$$

$$= \mathbb{E}_{i_k}\left[f(x_k) \right] - \alpha_k \mathbb{E}_{i_k}\left[\nabla f(x_k)^T \nabla f_{i_k}(x_k) \right] + \frac{L}{2} \alpha_k^2 \mathbb{E}_{i_k}\left[\| \nabla f_{i_k}(x_k) \|^2 \right]$$

independent of i_k by assumption

(Here we assume that α_k is independent of i_k, e.g. $\alpha_k = \frac{1}{L}$)

x_k is completely determined by the values of $i_0, i_1, \dots, i_{k-1}$

$$\Rightarrow \quad \mathbb{E}_{i_k}\left[f(x_k)\right] = f(x_k)$$

$$\Rightarrow \quad \mathbb{E}_{i_k}\left[\nabla f(x_k)^T u\right] = \nabla f(x_k)^T \mathbb{E}_{i_k}[u] \quad \forall u \in \mathbb{R}^d$$

Overall, we obtain

$$(3) \quad \mathbb{E}_{i_k}\left[f\left(x_k - \alpha_k \nabla f_{i_k}(x_k)\right)\right] \leq f(x_k) - \alpha_k \nabla f(x_k)^T \mathbb{E}_{i_k}\left[\nabla f_{i_k}(x_k)\right] + \frac{L\alpha_k^2}{2} \mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2\right]$$

- counterpart of the GD inequality
- used to prove $\mathbb{E}_{i_k}\left[f(x_{k+1})\right] \leq f(x_k)$ under the appropriate assumptions

$\Rightarrow$ Descent (decrease in f) on average

Assumptions on $\nabla f_{i_k}(x_k)$

At every iteration k,

a) $\mathbb{E}_{i_k}\left[\nabla f_{i_k}(x_k)\right] = \nabla f(x_k)$

"On average, the stochastic gradient behaves like the true gradient"

b) $\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2\right] \leq \|\nabla f(x_k)\|^2 + \sigma^2$ with $\sigma \geq 0$, $\sigma < \infty$

under a) $\Leftarrow$ $\mathbb{E}_{i_k}\left[\|\nabla f_{i_k}(x_k)\|^2 - \|\mathbb{E}_{i_k}[\nabla f_{i_k}(x_k)]\|^2\right] \leq \sigma^2 \rightarrow$ Noise/Variance level

"The norm of the stochastic gradient does not vary too much from the true gradient norm"

Using Assumptions (a) and (b) in (3), we obtain

$$\mathbb{E}_{i_k}\left[f\left(x_k - \alpha_k \nabla f_{i_k}(x_k)\right)\right] \leq \underbrace{f(x_k) - \alpha_k \|\nabla f(x_k)\|^2}_{\text{what we had for gradient descent}} + \frac{L\alpha_k^2}{2}\|\nabla f(x_k)\|^2 + \underbrace{\frac{L\alpha_k^2}{2}\sigma^2}_{\text{accounts for the uncertainty in choosing } i_k}$$