Optimization for Machine Learning

Part III : Stochastic Gradient Descent

Lionel Tondji

African Master's in Machine Intelligence

July 29, 2024

Final Project and Quiz

1. Lab on Stochastic Gradient Descent.
2. Group of max 3 students.
3. Write a small report containing answers to the questions and comments.
4. $\approx 15 + 5$ min presentation + QA per group on Monday.
5. Final Quiz on Friday.

The Training Problem

Solving the training problem :

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x^i), y^i) + \lambda R(w) = f(w) \tag{1}$$

A Datum function : $f_i(w) = \ell(h_w(x^i), y^i) + \lambda R(w)$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x^i), y^i) + \lambda R(w) &= \frac{1}{n} \sum_{i=1}^{n} (\ell(h_w(x^i), y^i) + \lambda R(w)) \\ &= \frac{1}{n} \sum_{i=1}^{n} f_i(w) \end{aligned}$$

Finite sum training problem :

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w) = f(w)$$

Reference Method: Gradient Method

$$\nabla\left(\frac{1}{n}\sum_{i=1}^{n} f_i(w)\right) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w)$$

Algorithm GD

starting points $w_0 \in \mathbb{R}^d$, **learning rate** $\alpha > 0$
for $k = 0, 1, 2, \cdots, T-1$ **do**
  $w_{k+1} = w_k - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(w_k)$
end for
Output w_T

Reference Method: Gradient Method

$$\nabla\left(\frac{1}{n}\sum_{i=1}^{n} f_i(w)\right) = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w)$$

Algorithm GD

starting points $w_0 \in \mathbb{R}^d$, **learning rate** $\alpha > 0$
for $k = 0, 1, 2, \cdots, T-1$ **do**
$w_{k+1} = w_k - \frac{\alpha}{n}\sum_{i=1}^{n} \nabla f_i(w_k)$
end for
Output w_T

Problem:

1. Each iteration requires computing a gradient $\nabla f_i(w)$ for each data point. One gradient for each cat on the internet!

Stochastic Gradient Descent

1. Is it possible to design a method that uses only the gradient of a single data function $f_i(w)$ at each iteration?

Stochastic Gradient Descent

1. Is it possible to design a method that uses only the gradient of a single data function $f_i(w)$ at each iteration?

Unbiased Estimate : Let j be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

Key Idea : Use

$$\nabla f_j(w) \approx \nabla f(w)$$

Exercise : Let $\sum_{i=1}^{n} p_i = 1$ and $j \sim p_j$. Show that

$$\mathbb{E}[\nabla f_j(w)/(np_j)] = \nabla f(w)$$

Stochastic Gradient Descent

Algorithm SGD : Constant step size

starting points $w_0 \in \mathbb{R}^d$, **learning rate** $\alpha > 0$
for $k = 0, 1, 2, \cdots, T-1$ **do**
  Sample $j \in \{1, \ldots, n\}$
  $w_{k+1} = w_k - \alpha \nabla f_i(w_k)$
end for
Output w_T

More reason why ML likes SGD

The training problem :

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell(h_w(x^i), y^i) + \lambda R(w) = f(w)$$

But we already know these labels.

The statistical learning problem: Minimize the expected loss over an unknown expectation

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(h_w(x), y) \right]$$

SGD can be applied to the statistical learning problem !

More reason why ML likes SGD

The statistical learning problem: Minimize the expected loss over an unknown expectation

$$\min_{w \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell(h_w(x), y)\right]$$

Algorithm SGD for learning

---

starting points $w_0 \in \mathbb{R}^d$, **learning rate** $\alpha_k > 0$

for $k = 0, 1, 2, \cdots, T-1$ **do**

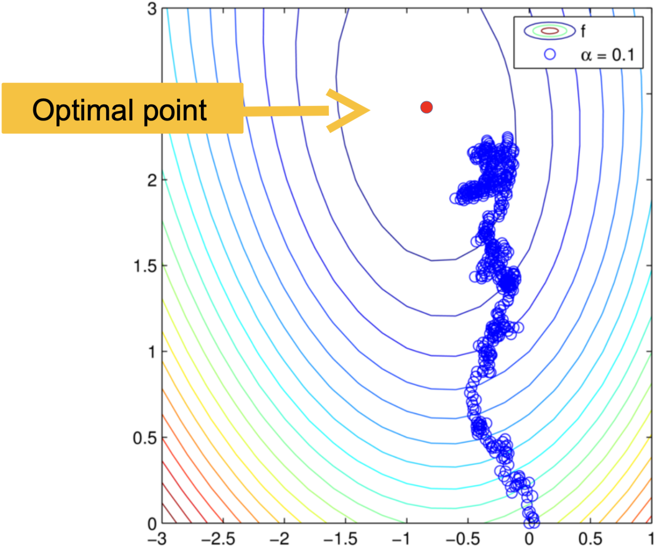
  Sample $(x, y) \sim \mathcal{D}$

  $w_{k+1} = w_k - \alpha_k \nabla \ell(h_{w_k}(x), y)$

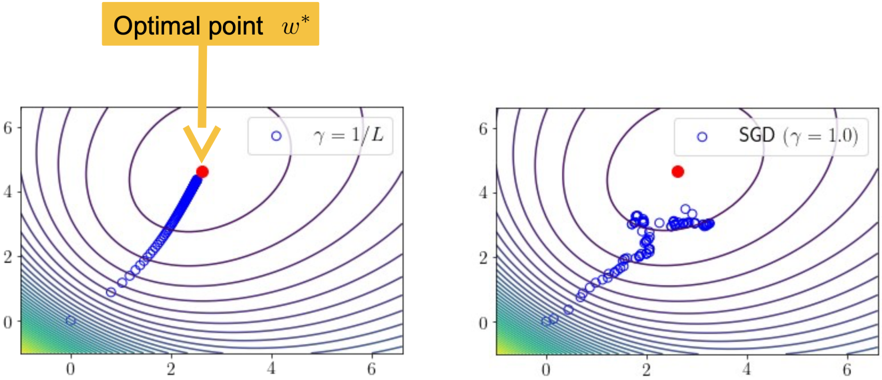
end for

Output $\bar{w}_T = \frac{1}{T}\sum_{k=1}^{T} w_k$

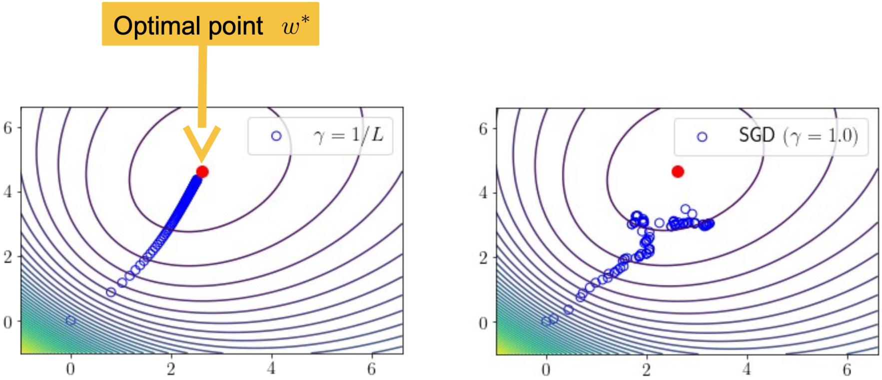
Stochastic Gradient Descent



GD vs Stochastic Gradient Descent



GD vs Stochastic Gradient Descent



Why does this happen? $\Rightarrow$ Need Assumptions

Assumptions for Convergence

1. **Strongly quasi-convexity**

$$f(w^*) \geq f(w) + \langle \nabla f(w), w^* - w \rangle + \frac{\mu}{2}\|w^* - w\|_2^2, \ \forall w$$

2. **Each f_i is convex and L_i smooth**

$$f_i(y) \leq f_i(w) + \langle \nabla f_i(w), y - w \rangle + \frac{L_i}{2}\|y - w\|_2^2, \ \forall w$$

$L_{\max} = \max_{i=1,\ldots,n} L_i$.

3. **Definition: Gradient Noise**

$$\sigma^2 := \mathbb{E}_j[\|\nabla f_j(w^*)\|_2^2]$$

Assumptions for Convergence

Example

Calculate the L_i's and $L_{\max}$ for

1. $f(w) = \frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$
2. $f(w) = \frac{1}{n}\sum_{i=1}^n ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}\|w\|_2^2$

Hint: A twice differentiable f_i is L_i -smooth if and only if

$$\nabla^2 f_i(w) \preccurlyeq L_i I \Leftrightarrow v^\top \nabla^2 f_i(w) v \leq L_i \|v\|_2^2, \ \forall v$$

1. $f(w) = \frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2$

$$f(w) = \frac{1}{2n}\|X^\top w - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}(x_i^\top w - y_i)^2 + \frac{\lambda}{2}\|w\|_2^2\right)$$
$$= \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

$$\nabla^2 f_i(w) = x_i x_i^\top + \lambda \preccurlyeq (\|x_i\|_2^2 + \lambda)I = L_i I$$
$$L_{\max} = \max_{i=1,\ldots,n}(\|x_i\|_2^2 + \lambda) = \max_{i=1,\ldots,n}(\|x_i\|_2^2) + \lambda$$

1. $f(w) = \frac{1}{n}\sum_{i=1}^{n} ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}\|w\|_2^2$

$$f_i(w) = ln(1 + e^{-y_i\langle w, a_i\rangle}) + \frac{\lambda}{2}\|w\|_2^2$$

$$\nabla f_i(w) = \frac{-y_i a_i e^{-y_i\langle w, a_i\rangle}}{1 + e^{-y_i\langle w, a_i\rangle}} + \lambda w$$

$$\begin{aligned}
\nabla^2 f_i(w) &= a_i a_i^\top \left(\frac{(1 + e^{-y_i\langle w, a_i\rangle})e^{-y_i\langle w, a_i\rangle}}{(1 + e^{-y_i\langle w, a_i\rangle})^2} - \frac{e^{-2y_i\langle w, a_i\rangle}}{(1 + e^{-y_i\langle w, a_i\rangle})^2} \right) + \lambda I \\
&= a_i a_i^\top \frac{e^{-y_i\langle w, a_i\rangle}}{(1 + e^{-y_i\langle w, a_i\rangle})^2} + \lambda I \\
&\preccurlyeq \left(\frac{\|a_i\|_2^2}{4} + \lambda \right) I = L_i I
\end{aligned}$$

since $\frac{e^x}{(1+e^x)^2} \leq \frac{1}{4}, \quad \forall x.$

Relationship between smoothness constants

Let $f(w)$ be convex.

1. Show that $f(w)$ is L-smooth with $L = \max_{w \in \mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w))$.
2. Thus $f_i(w)$ is L_i-smooth with $L_i = \max_{w \in \mathbb{R}^d} \lambda_{\max}(\nabla^2 f_i(w))$
3. Show that $L \leq \frac{1}{n} \sum_{i=1}^{n} L_i \leq L_{\max} := \max_{i=1,\ldots,n} L_i$

Relationship between smoothness constants

Let $f(w)$ be convex.

1. Show that $f(w)$ is L-smooth with $L = \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w))$.
2. Thus $f_i(w)$ is L_i-smooth with $L_i = \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f_i(w))$
3. Show that $L \leq \frac{1}{n}\sum_{i=1}^n L_i \leq L_{\max} := \max_{i=1,\ldots,n} L_i$

Proof.

From the Hessian definition of smoothness

$$\nabla^2 f(w) \preccurlyeq \lambda_{\max}(\nabla^2 f(w)) I \preccurlyeq \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w)) I$$

Relationship between smoothness constants

Let $f(w)$ be convex.

1. Show that $f(w)$ is L-smooth with $L = \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w))$.
2. Thus $f_i(w)$ is L_i-smooth with $L_i = \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f_i(w))$
3. Show that $L \leq \frac{1}{n}\sum_{i=1}^n L_i \leq L_{\max} := \max_{i=1,\ldots,n} L_i$

Proof.

From the Hessian definition of smoothness

$$\nabla^2 f(w) \preccurlyeq \lambda_{\max}(\nabla^2 f(w)) I \preccurlyeq \max_{w\in\mathbb{R}^d} \lambda_{\max}(\nabla^2 f(w)) I$$

Furthermore

$$\lambda_{\max}(\nabla^2 f(w)) = \lambda_{\max}(\frac{1}{n}\sum_{i=1}^n \nabla^2 f_i(w)) \leq \frac{1}{n}\sum_{i=1}^n \lambda_{\max}(\nabla^2 f_i(w)) \leq \frac{1}{n}\sum_{i=1}^n L_i$$

The final result now follows by taking the max over w, then max over i. $\square$

Complexity/Convergence

Theorem

If f is μ-strongly convex, f_i is convex and L_i-smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$, then the iterates of SGD satisfy

$$\mathbb{E}[\|w_k - w^*\|_2^2] \leq (1 - \alpha\mu)^k \|w_0 - w^*\|_2^2 + \frac{2\alpha}{\mu}\sigma^2 \tag{2}$$

Complexity/Convergence

Theorem

If f is μ-strongly convex, f_i is convex and L_i-smooth, $\alpha \in [0, \frac{1}{2L_{\max}}]$, then the iterates of SGD satisfy

$$\mathbb{E}[\|w_k - w^*\|_2^2] \leq (1 - \alpha\mu)^k \|w_0 - w^*\|_2^2 + \frac{2\alpha}{\mu}\sigma^2 \tag{2}$$

1. The first term shows that $\alpha \approx \frac{1}{\mu}$
2. The second term shows that $\alpha \approx 0$

Lemma

if $f_i : \mathbb{R}^n \to \mathbb{R} \cup \infty$ *convex and* $L_{\max}$*-smooth, then*

$$\mathbb{E}[\|\nabla f_j(w)\|_2^2] \leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Lemma

if $f_i : \mathbb{R}^n \to \mathbb{R} \cup \infty$ *convex and* $L_{\max}$*-smooth, then*

$$\mathbb{E}[\|\nabla f_j(w)\|_2^2] \leq 4L_{\max}(f(w) - f(w^*)) + 2\sigma^2$$

Proof.

Co-coercivity Lemma: If f convex and L-smooth

$$f(y) - f(x) \leq \langle \nabla f(y), y - x \rangle - \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_2^2$$

Applying this for f_i give us :

$$f_i(y) - f_i(x) \leq \langle \nabla f_i(y), y - x \rangle - \frac{1}{2L_{\max}}\|\nabla f_i(y) - \nabla f_i(x)\|_2^2$$

□

Proof

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(y)-\nabla f_i(x)\|_2^2 \leq 2L_{\max}\frac{1}{n}\sum_{i=1}^{n}(f_i(x)-f_i(y)+\langle\nabla f_i(y), y-x\rangle)$$
$$= 2L_{\max}(f(x)-f(y)+\langle\nabla f(y), y-x\rangle)$$

Take $y = x^* \in \mathit{arg}\min f(x)$, thus $\nabla f(x^*) = 0$ and

$$\frac{1}{n}\sum_{i=1}^{n}\|\nabla f_i(x^*)-\nabla f_i(x)\|_2^2 \leq 2L_{\max}(f(x)-f(x^*))$$

Using $\|\nabla f_i(x)\|_2^2 \leq 2\|\nabla f_i(x^*)-\nabla f_i(x)\|_2^2 + 2\|\nabla f_i(x^*)\|_2^2$

$$
\begin{aligned}
\mathbb{E}_j \|\nabla f_j(x)\|_2^2 &= \frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x)\|_2^2 \\
&\leq \frac{2}{n} \sum_{i=1}^{n} \|\nabla f_i(x^*) - \nabla f_i(x)\|_2^2 + 2\sigma^2 \\
&\leq 4L_{\max}(f(x) - f(x^*)) + 2\sigma^2
\end{aligned}
$$

Acknowledgements

Thanks for your attention!