STOCHASTIC GRADIENT DESCENT

Think about a problem in machine learning as the examples we saw.

Typically, one has a random variable X distributed according to some probability distribution (unknown). The model created depends on parameters $w \in \mathbb{R}^d$. The loss is, in fact, an expected loss:

$$\mathbb{E}(L(w; X))$$

This cannot be computed, so it is estimated. Data is observed, i.e., $\{x_i\}_{i=0,\dots,N-1}$. These are observations of the i.i.d. random variables $\{X_i\}_{i=0,\dots,N-1}$ (distributed as X).

Since

$$\frac{1}{N}\sum_{i=0}^{N-1} L(w; X_i)$$

is an unbiased estimator of

$$\mathbb{E}[L(w;X)]$$

we use the realization of this estimator on the data to estimate the expected loss:

$$\frac{1}{N}\sum_{i=0}^{N-1} L(w;x_i)$$

Let us call $L_i(w) = L(w;x_i)$.

Then the function to be optimized is of the form:

$$\frac{1}{N}\sum_{i=0}^{N-1} L_i(w)$$

Implementing a vanilla GD for this loss means:

$$w^{k+1} = w^k - \tau \frac{1}{N} \underbrace{\sum_{i=0}^{N-1} \nabla L_i(w^k)}$$

a lot of gradients must be computed, one for each element of the dataset.

This is computationally expensive

The idea behind Stochastic Gradient Descent (SGD) is to use only one of the function L_i, sampling randomly uniformly in the dataset. The algorithm generates a Markov chain of updated weights W^k as follows:

- W^0 is initialized randomly (e.g. uniformly or Gaussian)
- Assuming that we have observed $W^k = w^k$, we sample a random variable I_k uniformly in $\{0, \dots, N-1\}$ and we set

$$W^{k+1} = w^k - \tau \underbrace{\nabla L_{I_k}(w^k)}$$

the result is a random variable because this is a random variable.

We can prove a convergence result.

It is convenient to set

$$\hat{L}(w) := \frac{1}{N} \sum_{i=0}^{N-1} L_i(w)$$

<u>Remark</u>: If $0 < \lambda \leq D^2 L_i \leq \Lambda$ for $i = 0, \dots, N-1$, then

$$0 < \lambda \leq D^2 \hat{L} \leq \Lambda .$$

Indeed,

$$D^2 \hat{L}(w) = \frac{1}{N} \sum_{i=0}^{N-1} D^2 L_i(w)$$

$$\xi^T D^2 \hat{L}(w) \xi = \frac{1}{N} \sum_{i=0}^{N-1} \underbrace{\xi^T D^2 L_i(w) \xi}_{\text{this satisfies the inequality for } i=0,\dots,N-1.}$$

__Theorem__: Assume that

$$0 < \lambda \leq D^2 L_i \leq \Lambda \quad \text{for } i = 0, \dots, N-1,$$

Let $(W^K)_{K \in \mathbb{N}}$ the sequence of random variables generated by the SGD.
Let w^* be the unique minimum point of $\hat{L}$. Assume that $0 < \tau \leq \frac{1}{2\Lambda}$.
Then

$$\mathbb{E}\left[|W^K - w^*|^2\right] \leq (1 - \tau\lambda)^K \mathbb{E}\left[|W^0 - w^*|^2\right] + \frac{2\sigma^*}{\lambda}\tau$$

$$\left(\sigma^* = \mathrm{Var}\left[\nabla L_{I_K}(w^*)\right]\right)$$

__Proof__:

$$W^{K+1} = W^K - \tau \nabla L_{I_K}(W^K)$$

$$|W^{K+1} - w^*|^2 = |W^{K+1} - W^K + W^K - w^*|^2 =$$

$$= |W^{K+1} - W^K| + 2(W^{K+1} - W^K) \cdot (W^K - w^*) + |W^K - w^*|^2 =$$

$$= |W^K - w^*|^2 - 2\tau \nabla L_{I_K}(W^K) \cdot (W^K - w^*) + \tau^2 |\nabla L_{I_K}(W^K)|^2.$$

Take expectations:

$$\mathbb{E}\left[|W^{k+1}-w^*|^2\right] = \mathbb{E}\left[|W^k - w^*|^2\right] -$$

$$-2\tau\, \mathbb{E}\left[\nabla L_{I_k}(W^k)\cdot(W^k - w^*)\right] +$$

$$+\tau^2\, \mathbb{E}\left[|\nabla L_{I_k}(W^k)|^2\right].$$

Let us analyze the second term.

Assuming that $W^k = w^k$, we have

$$\mathbb{E}\left[\nabla L_{I_k}(W^k)\cdot(W^k - w^*) \mid W^k = w^k\right] =$$

$$= \mathbb{E}\left[\nabla L_{I_k}(w^k)\cdot(w^k - w^*)\right] =$$

$$= \frac{1}{N}\sum_{i=0}^{N-1} \nabla L_i(w^k)\cdot(w^k - w^*) =$$

$$= \nabla \hat{L}(w^k)\cdot(w^k - w^*)$$

Hence

$$\mathbb{E}\left[\nabla L_{I_k}(W^k)\cdot(W^k - w^*)\right] = \mathbb{E}\left[\nabla \hat{L}(W^k)\cdot(W^k - w^*)\right]$$

As we did in the case of vanilla GD,

$$\hat{L}(w^*) \geq \hat{L}(w^k) + \nabla\hat{L}(w^k)\cdot(w^* - w^k) + $$
$$+ \frac{\lambda}{2}|w^k - w^*|^2 \Rightarrow$$

$$\Rightarrow -2\tau\, \nabla\hat{L}(w^k)\cdot(w^k - w^*) \leq$$
$$\leq 2\tau\left(\hat{L}(w^*) - \hat{L}(w^k)\right) +$$
$$- \lambda\tau\, |w^k - w^*|^2$$

We have obtained:

$$\mathbb{E}\left[|W^{k+1} - w^*|^2\right] \leq (1-\lambda\tau)\,\mathbb{E}\left[|W^k - w^*|^2\right] +$$
$$+ 2\tau\, \mathbb{E}\left[\min\hat{L} - \hat{L}(W^k)\right]$$
$$+ \underbrace{\tau^2\, \mathbb{E}\left[|\nabla L_{I_k}(W^k)|^2\right]}$$

We want to absorb this positive term in the previous one.

Let us study it in detail.

$$\mathbb{E}\left[|\nabla L_{I_k}(W^k)|^2\right] \leq$$

$|a-b| \leq |a|+|b|$
$|a-b|^2 \leq |a|^2+|b|^2+2|a||b|$
$2|a||b| \leq |a|^2+|b|^2.$

$$\leq 2\,\mathbb{E}\left[|\nabla L_{I_k}(W^k) - \nabla L_{I_k}(w^*)|^2\right] +$$

$$+2\,\underbrace{\mathbb{E}\left[|\nabla L_{I_k}(w^*)|^2\right]}$$

This term is a number

$$\sigma^* = \mathbb{E}\left[|\nabla L_{I_k}(w^*)|^2\right] =$$

$$= \mathbb{E}\left[|\nabla L_{I_k}(w^*) - \nabla \hat{L}(w^*)|^2\right]$$

$$= \mathbb{E}\left[|\nabla L_{I_k}(w^*) - \mathbb{E}[\nabla L_{I_k}(w^*)]|^2\right]$$

$$= \mathrm{Var}\left[\nabla L_{I_k}(w^*)\right]$$

it's the gradient noise around the minimum.

As for the other term, we prove an inequality.

$$L_i(z) \leq L_i(w^k) + \nabla L_i(w^k)\cdot(z-w^k) + \\ + \frac{\Lambda}{2}|z-w^k|^2$$

$$L_i(z) \geq L_i(w^*) + \nabla L_i(w^*)\cdot(z-w^*) + \\ + \underbrace{\frac{\lambda}{2}|z-w^*|^2}_{\text{drop this}}$$

$\Rightarrow$

$$L_i(w^*) - L_i(w^k) \leq \nabla L_i(w^k)\cdot(z-w^k) + \\ + \nabla L_i(w^*)\cdot(w^*-z) + \frac{\Lambda}{2}|z-w^k|^2$$

Minimizing with respect to z:

$$\nabla L_i(w^k) - \nabla L_i(w^*) + \Lambda(z-w^k) = 0$$

$$z = w^k - \frac{1}{\Lambda}\left(\nabla L_i(w^k) - \nabla L_i(w^*)\right)$$

we obtain the tight bound

$$L_i(w^*) - L_i(w^k) \leq$$

$$\leq \nabla L_i(w^k)\cdot\left(-\frac{1}{\Lambda}\nabla L_i(w^k) + \frac{1}{\Lambda}\nabla L_i(w^*)\right) + \\ + \nabla L_i(w^*)\cdot\left(w^* - w^k + \frac{1}{\Lambda}\nabla L_i(w^k) - \frac{1}{\Lambda}\nabla L_i(w^*)\right) \\ + \frac{1}{2\Lambda}|\nabla L_i(w^k) - \nabla L_i(w^*)|^2 =$$

$$= -\frac{1}{2\Lambda} |\nabla L_i(w^*) - \nabla L_i(w^K)|^2 +$$

$$+ \nabla L_i(w^*) \cdot (w^* - w^K)$$

It follows that

$$\frac{1}{2\Lambda} |\nabla L_i(w^*) - \nabla L_i(w^K)|^2 \leq$$

$$\leq L_i(w^K) - L_i(w^*) + \nabla L_i(w^*) \cdot (w^* - w^K)$$

Applying this to L_{I_k} and taking the expectation:

$$\frac{1}{2\Lambda} \mathbb{E}\left[|\nabla L_{I_k}(w^*) - \nabla L_{I_k}(w^K)|^2 \right] \leq$$

$$\leq \hat{L}(w^K) - \min \hat{L} + \underbrace{\nabla \hat{L}(w^*)}_{=0\,!} \cdot (w^* - w^K)$$

Conditioning on $W^K = w^K$:

$$\frac{1}{2\Lambda} \mathbb{E}\left[|\nabla L_{I_k}(W^K) - \nabla L_{I_k}(w^*)|^2 \right] \leq$$

$$\leq \mathbb{E}\left[\hat{L}(W_k) - \min \hat{L} \right]$$

We insert this in the inequality we derived, estimating

$$\tau^2\, \mathbb{E}\left[|\nabla L_{I_k}(W^k)|^2\right] \leq$$

$$\leq 2\tau^2\, \mathbb{E}\left[|\nabla L_{I_k}(W^k) - \nabla L_{I_k}(w^*)|^2\right] +$$

$$+ 2\tau^2\, \mathbb{E}\left[|\nabla L_{I_k}(w^*)|^2\right] \leq$$

$$\leq 4\tau^2 \Lambda\, \mathbb{E}\left[\hat{L}(W_k) - \min \hat{L}\right] + 2\tau^2 \sigma^*$$

and thus

$$\mathbb{E}\left[|W^{k+1} - w^*|^2\right] \leq (1-\lambda\tau)\, \mathbb{E}\left[|W^k - w^*|^2\right] +$$

$$+ 2\tau\, \mathbb{E}\left[\min \hat{L} - \hat{L}(W^k)\right]$$

$$+ \tau^2\, \mathbb{E}\left[|\nabla L_{I_k}(W^k)|^2\right] \leq$$

$$\leq (1-\lambda\tau)\, \mathbb{E}\left[|W^k - w^*|^2\right] + 2\tau^2\sigma^* +$$

$$+ \underbrace{(2\tau - 4\tau^2\Lambda)}_{\text{if } \geq 0\text{; we can drop the whole term}}\, \mathbb{E}\left[\min \hat{L} - \hat{L}(W_k)\right]$$

$$1 - 2\tau\Lambda \geq 0 \Rightarrow \tau \leq \frac{1}{2\Lambda}$$

We have shown that

$$\mathbb{E}[|W^{k+1} - w^*|^2] \leq$$

$$\leq (1-\tau\lambda)\mathbb{E}[|W^k - w^*|^2] + 2\tau^2\sigma^* \leq$$

$$\leq (1-\tau\lambda)^2\mathbb{E}[|W^{k-1} - w^*|^2] + (1-\tau\lambda)2\tau^2\sigma^*$$

$$+ 2\tau^2\sigma^* \leq \dots \leq$$

$$\leq (1-\tau\lambda)^{k+1}\mathbb{E}[|W^0 - w^*|^2] +$$

$$+ \sum_{l=0}^{k} (1-\tau\lambda)^l 2\tau^2\sigma^* =$$

$$\leq (1-\tau\lambda)^{k+1}\mathbb{E}[|W^0 - w^*|^2] + \frac{1}{1-(1-\tau\lambda)} 2\tau^2\sigma^*$$

$\leq$ geometric series

$$= (1-\tau\lambda)^{k+1}\mathbb{E}[|W^0 - w^*|^2] + \frac{2\tau\sigma^*}{\lambda} \qquad \square$$

<u>Remark</u>: When implemented, for each epoch SGD shuffles the dataset and runs through the whole dataset. This is called Random Reshuffling.

MINI-BATCH GRADIENT DESCENT

This is the variant of SGD typically used in machine learning.

It is a compromise between the vanilla GD, which computes the gradient on the whole dataset, and SGD, which uses a single sample of the dataset.

We fix a batch size $b \in \mathbb{N}$, $b \geq 1$.
Given $B \subset \{0, 1, \dots, N-1\}$ such that $\# B = b$, we set:

$$L_B(w) := \frac{1}{\# B} \sum_{i \in B} L_i(w)$$

The Mini-batch SGD algorithm is:

- Initialize W^0 randomly (uniform or Gaussian distribution)

- Given that $W^k = w^k$, sample $B \subset \{0, 1, \dots, N-1\}$, $\# B = b$ uniformly among all possible subsets of size B and set $W^{k+1} = w^k - \tau \nabla L_B(w^k)$.

To prove convergence of mini-batch SGD, the steps are the same: we write the main differences.

<u>Step 1</u>: Show that

$$\mathbb{E}[|W^{k+1}-w^*|^2] = \mathbb{E}[|W^k-w^*|^2] -$$

$$-2\tau\ \mathbb{E}[\nabla L_B(W^k)\cdot(W^k-w^*)] +$$

$$+\tau^2\ \mathbb{E}[|\nabla L_B(W^k)|^2].$$

<u>Step 2</u>:

For the second term:

$$\mathbb{E}[\nabla L_B(W^k) \mid W^k = w^k] = \mathbb{E}[\nabla L_B(w^k)] =$$

$$= \sum_{\substack{B\subset\{0,\dots,N-1\}\\ \#B=b}} \frac{1}{\#B}\sum_{i\in B}\nabla L_i(w^k)\cdot\frac{1}{\binom{N}{b}} =$$

$$= \sum_{i=0}^{N-1}\ \sum_{\substack{B\subset\{0,\dots,N-1\}\\ \#B=b,\ i\in B}} \frac{1}{\#B}\cdot\frac{1}{\binom{N}{b}}\cdot\nabla L_i(w^k)$$

$$= \sum_{i=0}^{N-1}\binom{N-1}{b-1}\frac{1}{b}\,\frac{1}{\binom{N}{b}}\,\nabla L_i(w^k) =$$

$$= \sum_{i=0}^{N-1} \frac{(N-1)!}{(b-1)!(N-b)!} \cdot \frac{1}{b} \cdot \frac{b!(N-b)!}{N!} \nabla L_i(w^k)$$

$$= \frac{1}{N} \sum_{i=0}^{N-1} \nabla L_i(w^k) = \nabla \hat{L}(w^k).$$

<u>Step 3</u>: For the third term

Define

$$\sigma_b^* := \mathrm{Var}\left[\nabla L_B(w^*)\right]$$

and estimate

$$\mathbb{E}\left[|\nabla L_B(W^k)|^2\right] \le$$

$$\le 2\,\mathbb{E}\left[|\nabla L_B(W^k) - \nabla L_B(w^*)|^2\right] +$$

$$+ 2\underbrace{\mathbb{E}\left[|\nabla L_B(w^*)|^2\right]}_{\sigma_b^*}$$

<u>Step 4</u>: Estimate this using that

$$0 < \lambda \le D^2 L_B \le \Lambda.$$

The final result is:

$$\mathbb{E}\left[|W^k - w^*|^2\right] \leq (1-\tau\lambda)^k \,\mathbb{E}\left[|W^0 - w^*|^2\right] + \frac{2\sigma_b^*}{\lambda}\tau$$