Machine Learning and Neural Networks (MATH3431) Epiphany term, 2023

Handout 2: Stochastic gradient descent

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

Aim. To introduce the stochastic gradient descent (motivation, description, practical tricks, analysis in the convex scenario, and implementation).

Reading list & references:

- Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
- Bottou, L. (2012). Stochastic gradient descent tricks. In Neural networks: Tricks of the trade (pp. 421-436). Springer, Berlin, Heidelberg.

This is under development, it is subject to minor changes according to the Lecture, and it will be finalized around 1 day after the Lecture. It is given as guide before the lecture.

1. Motivations for stochastic gradient descent

Problem 1. Consider a learning problem $(\mathcal{H}, \mathcal{Z}, \ell)$. Learning may involve the computation of the minimizer $w^* \in \mathcal{H}$, where $\mathcal{H}$ is a class of hypotheses, of the risk function (RF) $R(w) = \mathrm{E}_{z\sim g}(\ell(w, z))$ given an unknown data generating model $g(\cdot)$ and using a known tractable loss $\ell(\cdot, \cdot)$; that is

$$w^* = \arg\min_{\forall w\in\mathcal{H}} (R_g(w)) = \arg\min_{\forall w\in\mathcal{H}} (\mathrm{E}_{z\sim g}(\ell(w, z))) \tag{1.1}$$

Remark 2. Gradient descent (GD) cannot be directly utilized to address Problem 1 (i.e., minimize the Risk function) because g is unknown, and because (1.1) involves an integral which may be computationally intractable. Instead it aims to minimize the ERF $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell(h, z_i)$ which ideally is used as a proxy when data size n is big (big-data).

Remark 3. The implementation of GD may be computationally impractical even in problems where we need to minimize an ERF $\hat{R}_n(w)$ if we have big data ($n \approx$big), This is because GD requires the recursive computation of the exact gradient $\nabla\hat{R}_n(w) = \frac{1}{n}\sum_{i=1}^{n} \nabla\ell(w, z_i)$ using all the data $\{z_i\}$ at each iteration. That may be too slow.

Remark 4. Stochastic gradient descent (SGD) aims at solving (1.1), and overcoming the issues in Remarks 2 & 3 by using an unbiased estimate of the gradient (required by GD) based on a sample properly drawn from g.

2. Stochastic gradient descent

2.1. Description.

Notation 5. For the sake of notation simplicity and generalization, we will present Stochastic Gradient Descent (SGD) in the following minimization problem

$$w^{*} = \arg \min_{\forall w \in \mathcal{H}} \left(f\left(w \right) \right) \tag{2.1}$$

where here $f : \mathbb{R}^{d} \rightarrow \mathbb{R}$, and $w \in \mathcal{H} \subseteq \mathbb{R}^{d}$; $f\left(\cdot\right)$ is the unknown function to be minimized, e.g., $f\left(\cdot\right)$ can be the risk function $R_{g}\left(w\right) = \mathrm{E}_{z \sim g}\left(\ell\left(w, z\right)\right)$.

Algorithm 6. *Stochastic Gradient Descent (SGD) with learning rate $\eta_t > 0$ for the solution of the minimization problem (**??**). where $\partial f\left(w^{(t)}\right)$ is the set of sub-gradients of f at $w^{(t)}$.*

For $t = 1, 2, 3, ...$ iterate:

(1) *compute*

$$w^{(t+1)} = w^{(t)} - \eta_{t} v_{t}, \tag{2.2}$$

where v_t is a random vector such that $E\left(v_t | w^{(t)}\right) \in \partial f\left(w^{(t)}\right)$

(2) *terminate if a termination criterion is satisfied, e.g.*

$$\textit{If } t \geq T \textit{ then STOP}$$

Remark 7. If f is differentiable at $w^{(t)}$, it is $\partial f\left(w^{(t)}\right) = \left\{\nabla f\left(w^{(t)}\right)\right\}$. Hence v_t is such as $\mathrm{E}\left(v_t | w^{(t)}\right) = \nabla f\left(w^{(t)}\right)$ in Algorithm 6 step 1.

Note 8. Assume $f$is differentiable (for simplicity). We can re-write (2.2) in the SGD Algorithm 6 as

$$w^{(t+1)} = w^{(t)} - \eta_{t} \left[\nabla f\left(w^{(t)} \right) + \xi_{t} \right], \tag{2.3}$$

where

$$\xi_{t} := v_{t} - \nabla f\left(w^{(t)} \right)$$

represents the (observed) noise introduced in (2.2) by using a random realization of the exact gradient.

Remark 9. Given T SGD algorithm iterations, the output of SGD can be (but not a exclusively),

(1) the average (after discarding the first few iterations of $w^{(t)}$ for stability reasons)

$$w_{\mathrm{SGD}}^{(T)} = \frac{1}{T} \sum_{t=1}^{T} w^{(t)} \tag{2.4}$$

(2) or the best value discovered

$$w_{\mathrm{SGD}}^{(T)} = \arg \min_{\forall w_t} \left(f\left(w^{(t)} \right) \right)$$

(3) or the last value discovered

$$w_{\mathrm{SGD}}^{(T)} = w^{(T)}$$

Note 10. SGD output converges to a local minimum, $w_{\mathrm{SGD}}^{(T)} \rightarrow w_{*}$ (in some sense), under different sets of regularity conditions. Section 2.2 has a brief analysis. To achieve this, Conditions 11 on the learning rate are inevitable and should be satisfied.

Condition 11. Regarding the learning rate (or gain) $\{\eta_t\}$ should satisfy conditions

(1) $\eta_t \geq 0$,
(2) $\sum_{t=1}^{\infty} \eta_t = \infty$
(3) $\sum_{t=1}^{\infty} \eta_t^2 < \infty$

Remark 12. The popular learning rates $\{\eta_t\}$ in Remark **??** in Handout 2, satisfy Condition 11.

Remark 13. Intuition on Condition 11. Assume that v_t is bounded. Condition 11((3)) aims at reducing the effect of the stochasticity in v_t (introduced noise ξ_t) because it implies $\eta_t \searrow 0$ as $t \to \infty$ and hence allows the chain to converge as

$$w^{(t+1)} - w^{(t)} = -\eta_t v_t \to 0.$$

Condition 11(2) prevents η_t from reducing too fast and allows the generated chain $\left\{w^{(t)}\right\}$ to be able to converge. E.g., after t iterations

$$\begin{aligned}\left\|w^{(t)} - w^*\right\| &= \left\|w^{(t)} \pm w^{(0)} - w^*\right\| \geq \left\|w^{(0)} - w^*\right\| - \left\|w^{(t)} - w^{(0)}\right\| \\ &\geq \left\|w^{(0)} - w^*\right\| - \sum_{t=0}^{\infty} \left\|w^{(t+1)} - w^{(t)}\right\| = \left\|w^{(0)} - w^*\right\| - \sum_{t=0}^{T-1} \|\eta_t v_t\|\end{aligned}$$

However if it was $\sum_{t=1}^{\infty} \eta_t < \infty$ it would be $\sum_{t=0}^{\infty} \|\eta_t v_t\| < \infty$ and hence $w^{(t)}$ will never converge to w^* if the seed $w^{(0)}$ is far enough from w^*.

2.2. **Analysis of SGD.**

Note 14. Recall that the stochastically comes from the random sub-gradient v_t, hence the expectations are under these random vectors.

Theorem 15. *Let $f(\cdot)$ be a convex and Lipschitz function. If we run SGD algorithm of f with learning rate $\eta_t > 0$ for T steps the output $w_{\text{GD}}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$* satisfies

$$\mathrm{E}\left(f\left(w_{\text{GD}}^{(T)}\right)\right) - f(w^*) \leq \frac{\mathrm{E}\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathrm{E}\|v_t\|^2 \tag{2.5}$$

Proof. Let $v_{1:t} = (v_1, ..., v_t)$. By Jensens' inequality (or see (**??**) in Handout 2)

$$\mathrm{E}\left(f\left(w_{\text{GD}}^{(T)}\right) - f(w^*)\right) \leq \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left(f\left(w^{(t)}\right) - f(w^*)\right)\right) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(f(w_t) - f(w^*)\right) \tag{2.6}$$

I will try to use Lemma **??** from Handout 2, hence I need to show

$$\mathrm{E}\left(f\left(w^{(t)}\right) - f(w^*)\right) \leq \mathrm{E}\left(\langle w^{(t)} - w^*, v_t\rangle\right) \tag{2.7}$$

where the expectation is under $v_{1:T}$. It is

$$\begin{aligned}\mathrm{E}_{v_{1:T}}\left(\langle w^{(t)} - w^*, v_t\rangle\right) =& \mathrm{E}_{v_{1:t}}\left(\langle w^{(t)} - w^*, v_t\rangle\right) \\ =& \mathrm{E}_{v_{1:t-1}}\left(\mathrm{E}_{v_{1:t}}\left(\langle w^{(t)} - w^*, v_t\rangle | v_{1:t-1}\right)\right) \qquad (\text{ law of total expectation })\end{aligned}$$

But $w^{(t)}$ is fully determined by $v_{1:t-1}$, (see (2.2)) so

$$\mathrm{E}_{v_{1:t-1}}\left(\mathrm{E}_{v_{1:t}}\left(\langle w^{(t)}-w^{*},v_{t}\rangle|v_{1:t-1}\right)\right)=\mathrm{E}_{v_{1:t-1}}\left(\langle w^{(t)}-w^{*},\mathrm{E}_{v_{1:t}}\left(v_{t}|v_{1:t-1}\right)\rangle\right)$$

As $w^{(t)}$ is fully determined by $v_{1:t-1}$ then $\mathrm{E}_{v_{1:t}}\left(v_{t}|v_{1:t-1}\right)=\mathrm{E}_{v_{1:t}}\left(v_{t}|w^{(t)}\right)\in\partial f\left(w^{(t)}\right)$, hence $\mathrm{E}_{v_{1:t}}\left(v_{t}|v_{1:t-1}\right)$ is a sub-gradient. By sub-gradient definition

$$\begin{aligned}\mathrm{E}_{v_{1:t-1}}\left(\langle w^{(t)}-w^{*},\mathrm{E}_{v_{1:t}}\left(v_{t}|v_{1:t-1}\right)\rangle\right)\geq&\mathrm{E}_{v_{1:t-1}}\left(f\left(w^{(t)}\right)-f\left(w^{*}\right)\right)\\=&\mathrm{E}_{v_{1:T}}\left(f\left(w^{(t)}\right)-f\left(w^{*}\right)\right)\end{aligned}\tag{2.8}$$

Hence combining (2.8), (2.7), and (2.7)

$$\mathrm{E}\left(f\left(w_{\mathrm{GD}}^{(T)}\right)-f\left(w^{*}\right)\right)\leq\frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(\langle w^{(t)}-w^{*},v_{t}\rangle\right)$$

Lemma **??** from Handout 2

$$\mathrm{E}\left(f\left(w_{\mathrm{GD}}^{(T)}\right)-f\left(w^{*}\right)\right)\leq\frac{\mathrm{E}\left\|w^{*}\right\|^{2}}{2\eta}+\frac{\eta}{2}\sum_{t=1}^{T}\mathrm{E}\left\|v_{t}\right\|^{2}$$

□

Remark 16. Note that the upper bound in (2.5) depends on the variation of $\mathrm{E}\left\|w^{*}\right\|^{2}$ and $\mathrm{E}\left\|v_{t}\right\|^{2}$ explicitely.

Proposition 17. *Let $f(\cdot)$ be a convex and Lipschitz function, and let $\mathcal{H}=\{w\in\mathbb{R}:\|w\|\leq B\}$. Assume we run SGD algorithm of $f(\cdot)$ with learning rate $\eta_{t}=\sqrt{\frac{B^{2}}{\rho^{2}T}}$ for T steps, and output $w_{\mathrm{SGD}}=\frac{1}{T}\sum_{t=1}^{T}w^{(t)}$. Then*

(1) *upper bound on the sub-optimality is*

$$\mathrm{E}\left(f\left(w_{\mathrm{GD}}^{(T)}\right)\right)-f\left(w^{*}\right)\leq\frac{B\rho}{\sqrt{T}}\tag{2.9}$$

(2) *a given level off accuracy ε such that $E\left(f\left(w_{GD}^{(T)}\right)\right)-f\left(w^{*}\right)\leq\varepsilon$ can be achieved after Titerations*

$$T\geq\frac{B^{2}\rho^{2}}{\varepsilon^{2}}.$$

Solution. Part 1 is a simple substitution from Proposition **??**, and part 2 is implied from part 1.

2.3. Implementation in the learning Problem 1.

Proposition 18. *Consider that ℓ is non-differentiable. For a randomly drawn $z\sim g(\cdot)$, the sub-gradient v of $\ell(w,z)$ at point w is an unbiased estimator of the sub-gradient of the risk $R_{g}(w)$ at point w.*

Proof. Let v be a sub-gradient of $\ell(w,z)$ at point w, then

$$\ell(u,z)-\ell(w,z)\geq\langle u-w,v\rangle\tag{2.10}$$

It is

$$R_g(u) - R_g(w) = \mathrm{E}_{z\sim g}\left(\ell(u,z) - \ell(w,z)\,|w\right) \geq \mathrm{E}_{z\sim g}\left(\langle u-w, v\rangle | w\right)$$
$$= \langle u-w, \mathrm{E}_{z\sim g}(v|w)\rangle$$

Hence the sub-gradient of $\ell(w,z)$ is an unbiased estimator of the sub-gradient of $R_g(w)$. □

Proposition 19. *Consider that ℓ is non-differntiable. If $v = \nabla_w \ell(w,z)$, then*

$$E_{z\sim g}(v|w) = \nabla_w R_g(w)$$

Algorithm 20. *Stochastic Gradient Descent (SGD) with learning rate $\eta_t > 0$ for Problem 1.*
For $t = 1, 2, 3, ...$ iterate:

(1) *sample $z^{(t)} \sim g$, and compute $v_t$such that $v_t \in \partial \ell\left(w^{(t)}, z^{(t)}\right)$*
(2) *compute*

$$w^{(t+1)} = w^{(t)} - \eta_t v_t, \tag{2.11}$$

(3) *terminate if a termination criterion is satisfied, e.g.*

$$\textit{If } t \geq T \textit{ then STOP}$$

Remark 21. In Algorithm 20 and step 1 sampling $z^{(t)} \sim g$ can be performed either by directly sampling from g (when possible), or by sub-sampling (with or without replacement) from the given dataset $\{z_i; i = 1, ..., n\}$.

2.4. **Variations.**

2.5. **Examples.**