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

Handout 3: 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.

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 $\hat{R}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(w, z_i)$ 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 estimator of the actual gradient (or some sub-gradient) 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 present Stochastic Gradient Descent (SGD) in the following minimization problem

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

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

Algorithm 6. *Stochastic Gradient Descent (SGD) with learning rate $\eta_t > 0$ for the solution of the minimization problem (2.1)*

For $t = 1, 2, 3, \ldots$ 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). To compare SGD with GD, 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, $w_{\mathrm{SGD}}^{(T)} \to w_*$ (in some sense), under different sets of regularity conditions. Section 4 has a brief analysis. To achieve this, Conditions 11 on the learning rate are rather 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 9 in Handout 2 satisfy Condition 11 and hence can be used in SGD too.

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 $t \to \infty$, which if it was not the case then

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

may not be satisfied and the chain may not converge. 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 $w^{(t)}$ would never converge to w^* if the seed $w^{(0)}$ is far enough from w^*.

Note 14. Following is a variation of SGD (Algorithm 6) to account for bounded cases such as $w \in \mathcal{H}$.

3. Stochastic gradient with projection

Remark 15. Consider the scenario in Problem 1 where the Risk function is non-convex in $\mathbb{R}^d$ but convex in the restricted hypothesis set $\mathcal{H}$ e.g. $\mathcal{H} = \{w : \|w\| \leq B\}$; hence the learning problem requires to discover w^* in the restricted/bounded set $\mathcal{H}$. Direct implementation of SGD Algorithm 6 may produce a chain stepping out $\mathcal{H}$ and hence an output $w_{\text{SGD}} \notin \mathcal{H}$. To address this issue, SGD can be modified to include a projection step guarantying $w \in \mathcal{H}$ as in Algorithm 16.

Algorithm 16. *Stochastic Gradient Descent with projection and with learning rate $\eta_t > 0$ for the solution of the minimization problem (2.1)*
For $t = 1, 2, 3, \ldots$ iterate:

(1) *compute*

$$w^{\left(t+\frac{1}{2}\right)} = w^{(t)} - \eta_t v_t, \tag{3.1}$$

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

(2) *compute*

$$w^{(t+1)} = \arg\min_{w \in \mathcal{H}} \left(\left\|w - w^{\left(t+\frac{1}{2}\right)}\right\|\right) \tag{3.2}$$

(3) *terminate if a termination criterion is satisfied*

4. Analysis of SGD (Algorithm 6)

Note 17. Recall that the stochasticity of SGD comes from the stochastic sub-gradient v_t; hence the expectations below are under these random vectors distributions.

Theorem 18. *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 $w_{\mathrm{GD}}^{(T)} = \frac{1}{T}\sum_{t=1}^{T} w^{(t)}$* satisfies

$$\mathrm{E}\left(f\left(w_{\mathrm{GD}}^{(T)}\right)\right) - f(w^*) \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\mathrm{E}\|v_t\|^2 \tag{4.1}$$

Proof. Let $v_{1:t} = (v_1, ..., v_t)$. By Jensens' inequality (or see (4.3) in Handout 2)

$$\mathrm{E}\left(f\left(w_{\mathrm{GD}}^{(T)}\right) - f(w^*)\right) \leq \mathrm{E}\left(\frac{1}{T}\sum_{t=1}^{T}\left(f\left(w^{(t)}\right) - f(w^*)\right)\right) = \frac{1}{T}\sum_{t=1}^{T}\mathrm{E}\left(f\left(w^{(t)}\right) - f(w^*)\right) \tag{4.2}$$

I will try to use Lemma 22 from Handout 2, hence I need to show

$$\mathrm{E}\left(f\left(w^{(t)}\right) - f(w^*)\right) \leq \mathrm{E}\left(\langle w^{(t)} - w^*, v_t\rangle\right) \tag{4.3}$$

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

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

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

$$\begin{aligned}\mathrm{E}_{v_{1:t-1}}\left(\langle w^{(t)} - w^*, \mathrm{E}_{v_{1:t}}(v_t|v_{1:t-1})\rangle\right) \geq& \mathrm{E}_{v_{1:t-1}}\left(f\left(w^{(t)}\right) - f(w^*)\right)\\ =& \mathrm{E}_{v_{1:T}}\left(f\left(w^{(t)}\right) - f(w^*)\right)\end{aligned} \tag{4.4}$$

Hence combining (4.4), (4.3), and (4.3)

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

Lemma 22 from Handout 2

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

□

Remark 19. Note that the upper bound in (4.1) depends on the variation of v_t as

$$\mathrm{E}\left\|v_t\right\|^2=\sum_{j=1}^{d}\operatorname{Var}\left(v_{t,j}\right)+\sum_{j=1}^{d}\left(\mathrm{E}\left(v_{t,j}\right)\right)^2 \tag{4.5}$$

where d is the dimension of $v_t=(v_{t,1},...,v_{t,d})$.

Proposition 20. *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 $w_{\mathrm{SGD}}^{(T)}=\frac{1}{T}\sum_{t=1}^{T}w^{(t)}$. Then*

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

$$\mathrm{E}\left(f\left(w_{\mathrm{SGD}}^{(T)}\right)\right)-f\left(w^*\right)\leq\frac{B\rho}{\sqrt{T}} \tag{4.6}$$

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

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

Solution. Part 1 is a simple substitution from Proposition 28, and part 2 is implied from part 1.