Deep Learning (SS2020)
Seminar 3

April 17, 2020

Assignment 1 (Sampling with Replacement)

a. Let the dataset contain n points. During an epoch, we draw a random point n times. What is the probability that point i has been drawn at least once? What is the limit of this probability as $n \to \infty$.

   Hint1: Write out the probability that a point has not been drawn in n trials.

   Hint2: To compute the limit use L'Hôpital's rule
   (or compute e.g. with `www.wolframalpha.com`)

b. See the "Coupon collector's problem" on wikipedia. What is the expected number of epochs we need to run to have each data point being drawn at least once?

Assignment 2 (EWA) Consider the running average $\mu_t = (1 - q_t)\mu_{t-1} + q_t X_t$ (SGD lecture slide 13).

a. Define a sequence q_t such that in the beginning the running average gives the equally weighted mean of the observations till the time t and in a longer run, it becomes equivalent to the exponentially weighted average.

b. What is the setting of q for the EWA, such that its smoothing effect is equivalent to a plain average of n points, as measured by the equal variance reduction?

   Hint: Assuming all observations have variance 1, the variance of EWA at step t is given by $\sum_{k=1}^{t} w_k^2$, where $w_k = (1-q)^{t-k} q$ for $k = 1, \ldots, t$. Find this sum using geometric series in the limit $t \to \infty$, i.e. when the initialization effect becomes unimportant. (The claim in the lecture that it is a constant value for all t was incorrect).

Assignment 3 (Momentum)

a. Consider SGD with momentum:

$$\begin{aligned} v_{t+1} &= \mu v_t + g_t \\ \theta_{t+1} &= \theta_t - \varepsilon v_{t+1}, \end{aligned} \tag{1}$$

where θ is the parameter vector we optimize, v is the velocity with momentum and g_t is the gradient at θ_t.
Express θ_{t+1} without using the velocity sequence, e.g., using only θ_t, θ_{t-1}, g_t and g_{t-1}.

b. Do the same for SGD with Nesterov momentum:

$$\begin{aligned} v_{t+1} &= \mu v_t + g_t \\ \theta_{t+1} &= \theta_t - \varepsilon\big(g_t + \mu v_{t+1}\big). \end{aligned} \tag{2}$$

Assignment 4 (CNNs)

a. Show that convolution is equivariant to sub-pixel translations of an image. A sub-pixel translation is implemented as a bilinear interpolation technique.

b. What is the size of the receptive field of one unit in the output of a fully convolutional network with layers without padding:
conv(5×5, stride 1, dilation 1)
conv(3×3, stride 1, dilation 2)
conv(3×3 stride 2, dilation 1),

where dilation 1 means standard convolution without holes and dilation 2 is as illustrated in the CNN lecture slide 23.