TTIC 31230 Fundamentals of Deep Learning

SGD Problems.

Problem 1. Variance of running averages. For two independent random variables x and y and a weighted sum $s = ax + by$ we have

$$\sigma_s^2 = a^2\sigma_x^2 + b^2\sigma_y^2$$

Now consider a runing average for computing $\hat{\mu}_1, \ldots, \hat{\mu}_t$ from $x_1, \ldots, x_t$

$$\hat{\mu}_0 = 0$$

$$\hat{\mu}_t = \left(1 - \frac{1}{N}\right)\hat{\mu}_{t-1} + \frac{1}{N}x_t$$

(a) Assume that the values of x_t are independent and identically distributed with variance σ_x^2. We now have that $\hat{\mu}_t$ is a random variable depending on the draws of x_t. The random variable $\hat{\mu}_t$ has a variance $\sigma_{\hat{\mu},t}^2$. Assume that as $t \to \infty$ we have that $\sigma_{\hat{\mu},t}^2$ converges to a limit (it does). Solve for this limit $\sigma_{\hat{\mu},\infty}^2$. Your solution should yield that for $N = 1$ we have $\sigma_{\hat{\mu},\infty}^2 = \sigma_x^2$ (a sanity check).

Solution: The limit must satisfy

$$\sigma_{\hat{\mu},\infty}^2 = \left(1 - \frac{1}{N}\right)^2 \sigma_{\hat{\mu},\infty}^2 + \frac{1}{N^2}\sigma_x^2$$

We can then solve for $\sigma_{\hat{\mu},\infty}^2$

$$\begin{aligned}
\sigma_{\hat{\mu},\infty}^2 &= \left(1 - \frac{2}{N} + \frac{1}{N^2}\right)\sigma_{\hat{\mu},\infty}^2 + \frac{1}{N^2}\sigma_x^2 \\
0 &= \left(\frac{-2}{N} + \frac{1}{N^2}\right)\sigma_{\hat{\mu},\infty}^2 + \frac{1}{N^2}\sigma_x^2 \\
&= \left((-2) + \frac{1}{N}\right)\sigma_{\hat{\mu},\infty}^2 + \frac{1}{N}\sigma_x^2 \\
\sigma_{\hat{\mu},\infty}^2 &= \frac{1}{\left(2 - \frac{1}{N}\right)N}\ \sigma_x^2
\end{aligned}$$

(b) Compare your answer to (a) with the variance of an average of N values of x_t defined by

$$\hat{\mu} = \frac{1}{N}\sum_{t=1}^{N} x_t$$

Solution: For an average of N we have $\sigma^2_{\hat{\mu}} = \sigma^2_x/N$. For N large we have that the answer to part (a) is about half as large.

Problem 2. Reformulating Momentum as a Running Average. Consider the following running update equation.

$$\begin{aligned} y_0 &= 0 \\ y_t &= \left(1 - \frac{1}{N}\right) y_{t-1} + x_t \end{aligned}$$

(a) Assume that y_t converges to a limit, i.e., that $\lim_{t\to\infty} y_t$ exists. If the input sequence is constant with $x_t = c$ for all $t \geq 1$, what is $\lim_{t\to\infty}\ y_t$? Give a derivation of your answer (Hint: you do not need to compute a closed form solution for y_t).

Solution:
The limit y_∞ must satisfy

$$y_\infty = \left(1 - \frac{1}{N}\right) y_\infty + c$$

giving $y_\infty = Nc$.

(b) y_t is a running average of what quantity?

Solution: The update can be rewritten as

$$y_t = \left(1 - \frac{1}{N}\right) y_{t-1} + \frac{1}{N}(Nx_t)$$

so y_t is the running average of Nx_t.

(c) Express y_t as a function of μ_t where μ_t is defined by

$$\begin{aligned} \mu_0 &= 0 \\ \mu_t &= \left(1 - \frac{1}{N}\right) \mu_{t-1} + \frac{1}{N} x_t \end{aligned}$$

Solution: y_t is the running average of Nx_t which equals N times the running average of x_t so we have

$$y_t = N\mu_t$$

Problem 3. Bias Correction Consider the following update equation for computing $y_1, \ldots, y_t$ from $x_1, \ldots, x_t$.

$$y_t \quad = \quad \left(1 - \frac{1}{\min(t, N)}\right) y_{t-1} + \frac{1}{\min(t, N)} \, x_t$$

If $x_t = c$ for all $t \geq 1$ give a closed form solution for y_t.

Solution: For $t = 1$ we get $y_1 = x_1 = c$. We then get that y_{t+1} is a convex combination of y_t and x_t which maintains the invariant that $y_t = c$.

Problem 4. Batch Size Coupling to RMSProp and Adam. Consider the following for-loop representation of a batch of matrix-vector products.

$$\text{for } b, i, j \;\; y[b, j] \mathrel{+}= W[j, i]x[b, i]$$

(a) Write the for-loop representation of back-propagation to W.grad following the convention that parameter gradients are averaged over the batch.

Solution:

$$\text{for } b, i, j \;\; w.\text{grad}[j, i] \mathrel{+}= \frac{1}{B} \, y.\text{grad}[b, j]x[b, i]$$

(b) Write a for-loop representation for computing W.grad$[b, i, j]$ where this is the derivative of loss with respect to $W[i, j]$ for batch element b.

Solution:

$$\text{for } b, i, j \;\; w.\text{grad}[b, j, i] \mathrel{+}= y.\text{grad}[b, j]x[b, i]$$

(c) Consider

$$W.\text{grad2}[j, i] = \frac{1}{B} \sum_b W.\text{grad}[b, j, i]^2$$

Is it possible to compute W.grad2$[j, i]$ from W.grad$[j, i]$? Explain your answer.

Solution: No. W.grad2$[j, i]$ is the average over the batch of the of the square of the gradient. The average value does not determine the average square value — the average value does not determine the variance.

(d) Explain how your answer to (c) is related to batch size scaling of RMSProp and Adam.

Solution: Adam and RMSProp both compute a running average of $\hat{g}[i]^2$ defined by

$$s_{t+1}[i] = \left(1 - \frac{1}{N_s}\right) s_t + \frac{1}{N_s} \hat{g}[i]^2$$

At batch sized greater than 1 this fails to take into account the variance of the gradiants within the batch. This implies that $s_t[i]$ will be reduced as the batch size increases and in the limit of large batches $s_t[i]$ will converge to the mean squared rather than the second moment.

Problem 5. This problem is on batch size scaling of continuous time stochastic differential equation (SDE) models of SGD. We consider batched SGD as defined by

$$\Phi \mathrel{-}= \eta \hat{g}^B$$

where $\hat{g}^B$ is the average of B sampled gradients. Let g be the average gradient $g = E\ \hat{g}$.

The covariance matrix at batch size B is

$$\Sigma^B[i,j] = E\ (\hat{g}^B[i] - g[i])(\hat{g}^B[j] - g[j]).$$

The continuous time stochastic differential equation model is

$$\Phi(t + \Delta t) = \Phi(t) - g\Delta t + \epsilon\sqrt{\Delta t} \quad \epsilon \sim \mathcal{N}(0, \eta\Sigma^B)$$

Show that for $\eta = B\eta_0$ the SDE is determined by η_0 independent of B.

Solution:

$$\begin{aligned}
\Sigma^B[i,j] &= E\ (\hat{g}^B[i] - g[i])(\hat{g}^B[j] - g[j]) \\
&= \frac{1}{B^2} E \left(\sum_b \hat{g}_b[i] - g[i]\right)\left(\sum_b \hat{g}_b[j] - g[j]\right) \\
&= \frac{1}{B^2} E \sum_{b,b'} (\hat{g}_b[i] - g[i])\,(\hat{g}_{b'}[j] - g[j]) \\
&= \frac{1}{B^2} \sum_{b,b'}\ E\ \ (\hat{g}_b[i] - g[i])\,(\hat{g}_{b'}[j] - g[j]) \\
&= \frac{1}{B^2} \sum_{b}\ E\ \ (\hat{g}_b[i] - g[i])\,(\hat{g}_{b}[j] - g[j]) \\
&= \frac{1}{B} \Sigma^1[i,j]
\end{aligned}$$

So for $\eta = B\eta_0$ we have $\eta\Sigma^B = \eta_0\Sigma^1$ which yields the equivalence.