TTIC 31230 Fundamentals of Deep Learning

SGD Problems.

Problem 1. 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) If the input sequence is constant, i.e., if $x_t = c$ 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).

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

(c) Consider the following running average of x_t.

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

Express y_t as a function of μ_t.

Problem 2. Consider the following update equation

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

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

Problem 3. This problem is on batch size scaling of RMSProp. Consider the following for-loop representation of a batch of matrix-vector products.

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

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

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

(c) Consider

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

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

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

Problem 4. This problem is on batch size scaling of Langevin dynamics. We consider batched SGD as defined by

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

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

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

Langevin dynamics 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 $\eta = B\eta'$ the Langevin dynamics is determined by η' independent of B.

Solution:

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

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