TTIC 31230 Fundamentals of Deep Learning

SGD Problems.

Problem 1. The stationary distribution with minibatching. This problem is on batch size scaling and stationary distributions (temperature). 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 stochastic differential equation model of SGD 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 stationary distribution is determined by η_0 independent of B.

Solution:

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

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

Problem 2. The stationary distribution of weighted average updates. In this problem we consider the more general principle that the stationary distribution (the temperature) is determined by the effect of each individual training point on the parameter vector. This principle was used in the claim that in the presence of momentum setting the learning rate by $\eta = (1-\mu)B\eta_0$ yields a temperature determined by η_0 independent of μ and B. In this problem we justify the general principle of examining the influence of each individual training point.

Let $\hat{g}_1, \ldots, \hat{g}_N$ be the loss gradients of N individual training points (Batch size 1). Consider a weighted sum (such as that used in momentum).

$$\Delta\Phi = \sum_i \alpha_i \hat{g}_i$$

Assume the updates are small so that for any given training point i we have that $\hat{g}_i$ is uneffected by the drift in the parameter vector. In that case, even if parameter updates are being made between gradient measurements, the random variables g_i are essentially independent and identically distributed (over the random draw of a training point). Let Σ_g be the covariance matrix of the distribution of the random variable $\hat{g}_i$. Let $\Sigma_{\Delta\Phi}$ be the covariance matrix of the random variable $\Delta\Phi$. Show

$$\Sigma_{\Delta\Phi} = \left(\sum_i \alpha_i^2\right) \Sigma_g$$