TTIC 31230 Fundamentals of Deep Learning

Generalization Problems

Problem 1. Consider the regularized objective

$$\Phi^* = \underset{\Phi}{\text{argmin}} \; E_{(x,y)\sim\text{Train}} \left(\mathcal{L}(\Phi, x, y) + \frac{\lambda}{2N} ||\Phi||^2 \right)$$

By setting the gradient of the objective to zero, solve for the average gradient g as a function of Φ^*.

Problem 2. Consider any probability distribution $P(h)$ over an discrete class $\mathcal{H}$. Assume $0 \leq \mathcal{L}(h, x, y) \leq L_{\max}$. Define

$$\mathcal{L}(h) = E_{(x,y)\sim\text{Pop}} \, \mathcal{L}(h, x, y)$$

$$\hat{\mathcal{L}}(h) = E_{(x,y)\sim\text{Train}} \, \mathcal{L}(h, x, y)$$

We now have the theorem that with probability at least $1 - \delta$ over the draw of training data the following holds simultaneously for all h.

$$\mathcal{L}(h) \leq \frac{10}{9} \left(\hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N} \left(\ln \frac{1}{P(h)} + \ln \frac{1}{\delta} \right) \right) \quad (1)$$

This motivates

$$h^* = \underset{h}{\text{argmin}} \; \hat{\mathcal{L}}(h) + \frac{5L_{\max}}{N} \ln \frac{1}{P(h)} \quad (2)$$

The Bayesian maximum a-posteriori (MAP) rule is

$$h^* = \underset{h}{\text{argmax}} \; P(h) \prod_{(x,y)\in\text{Train}} P(y|x, h) \quad (3)$$

For $\mathcal{L}(h, x, y) = -\ln P(y|x, h)$ (cross entropy loss) rewrite (2) so as to be as similar to (3) as possible. Note that (1) holds independent of any "truth" of the "prior" P.

Problem 3.

(a) Consider a model where the parameter vector Φ has d parameters each of which is represented by a 16 bit floating point number. Express the bound (1) in problem 2 in terms of the dimension d assuming all parameter vectors are equally likely.

(b) Define a probability distribution over variable precision floating point representations where any number of bits of mantissa is possible and any number of bits of exponent is possible and then express the bound (1) in terms of this variable-precision representation of numbers.