1 L_1 and L_2 Regularization

1.1 Learning Objectives

1. Explain the concept of a sequence of nested hypothesis spaces, and explain how a complexity measure (of a function) can be used to create such a sequence.

2. Given a base hypothesis space of decision functions (e.g. affine functions), a performance measure for a decision function (e.g. empirical risk on a training set), and a function complexity measure (e.g. Lipschitz continuity constant of decision function), give the corresponding optimization problem in Tikhonov and Ivanov forms.

3. For some situations (i.e. combinations of base hypothesis space, performance measure, and complexity measure), we claimed that Tikhonov and Ivanov forms are equivalent. Be able to explain what this means and write it down mathematically.

4. In particular, the Tikhonov and Ivanov formulations are equivalent for lasso and ridge regression. Be comfortable switching between the formulations to assist with interpretations (e.g. the classic L1 regularization picture with the norm ball is based on the Ivanov formulation).

1.2 Concept Check Questions

1. Consider the following two minimization problems:

$$\arg\min_w \Omega(w) + \frac{\lambda}{n}\sum_{i=1}^{n} L(f_w(x_i), y_i)$$

and

$$\arg\min_w C\Omega(w) + \frac{1}{n}\sum_{i=1}^{n} L(f_w(x_i), y_i),$$

where $\Omega(w)$ is the penalty function (for regularization) and L is the loss function. Give sufficient conditions under which these two give the same minimizer.

2. ($\star$) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. Prove that $\|\nabla f(x)\|_2 \leq L$ if and only if f is Lipschitz with constant L.

3. ($\star$) Let $\hat{w}$ denote the minimizer for

$$\begin{array}{ll} \text{minimize}_w & \|Xw - y\|_2^2 \\ \text{subject to} & \|w\|_1 \leq r. \end{array}$$

Prove that $f(x) = \hat{w}^T x$ is Lipschitz with constant r.

4. Two of the plots in the lecture slides use the fact that $\|\hat{w}\|/\|\tilde{w}\|$ is always between 0 and 1. Here $\hat{w}$ is the parameter vector of the linear model resulting from the regularized least squares problem. Analgously, $\tilde{w}$ is the parameter vector from the unregularized problem. Why is this true that the quotient lies in $[0, 1]$?

5. Explain why feature normalization is important if you are using L_1 or L_2 regularization.