1 L_1 and L_2 Regularization

1.1 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.

Solution. Let $C = 1/\lambda$. Then the two objectives differ by a constant factor.

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

Solution. First suppose $\|\nabla f(x)\|_2 \le L$ for some $L \ge 0$ and all $x \in \mathbb{R}^n$. By the mean value theorem we have, for any $x, y \in \mathbb{R}^n$,

$$f(y) - f(x) = \nabla f(x + \xi(y-x))^T(y-x),$$

where ξ is some value between 0 and 1. Taking absolute values on each side we have

$$|f(y) - f(x)| = |\nabla f(x + \xi(y-x))^T(y-x)| \le \|\nabla f(x + \xi(y-x))\|_2\|y-x\|_2$$

by Cauchy-Schwarz. Applying our bound on the gradient norm proves f is Lipschitz with constant L.

Conversely, suppose f is Lipschitz with constant L. Note that

$$|\nabla f(x)^T v| = |f'(x; v)| = \left|\lim_{t\to 0}\frac{f(x+tv) - f(x)}{t}\right| \le \lim_{t\to 0}\frac{|t|L\|v\|}{|t|} = L\|v\|.$$

Letting $v = \nabla f(x)$ we obtain $\|\nabla f(x)\|_2^2 \le L\|\nabla f(x)\|_2$ giving the result.

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

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

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

Solution. Note that $\|w\|_2 \le \|w\|_1 \le r$, so the argument from class gives the result. To see the inequality, note that

$$\|w\|_1^2 = (|w_1| + \cdots + |w_n|)^2 \ge |w_1|^2 + \cdots + |w_n|^2 = \|w\|_2^2.$$

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 $[0, 1]$?

Solution. We assume Ivanov regularization (since Tikhonov is equivalent). We know that

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{w}^T x_i - y_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n}(\hat{w}^T x_i - y_i)^2$$

since $\tilde{w}$ is the solution to the unconstrained minimization. But if $\|\tilde{w}\| \leq \|\hat{w}\|$ then $\|\tilde{w}\|$ is feasible for the regularized problem, so $\|\hat{w}\| = \|\tilde{w}\|$. Thus $\|\tilde{w}\| \geq \|\hat{w}\|$.

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

   Solution. Suppose you have a model $y = w^T x$ where x_1 is a very correlated with y, but the feature is measured in meters. Thus $w_1 = 4$ would mean each increase in x_1 by 1 meter yields an increase in y by 4. Now suppose we change the units of w_1 to kilometers by scaling it. This would require us to change w_1 to 4000 to achieve the same decision function. While this has no effect on the loss $(y - w^T x)^2$ it has a significant effect on $\lambda\|w\|_2^2$ or $\lambda\|w\|_1$. For example, even if $x_2, \ldots, x_n$ had very little relationship with y, we would still undervalue w_1 due to the regularization.