1 Introduction

The goal is to directly optimize for the optimal regularization strength $\hat{\lambda}$ and avoid the need for cross validation. Let $\mathcal{Z}$ be a dataset of i.i.d. data points. We break up $\mathcal{Z}$ into a training set $\mathcal{Z}_t$ and validation set $\mathcal{Z}_v$ such that $\mathcal{Z}_t \cup \mathcal{Z}_v = \mathcal{Z}$. We then use regularized loss minimization to estimate a parameter vector

$$\hat{\mathbf{w}}_\lambda = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{\mathbf{z}\in\mathcal{Z}_t} \ell(\mathbf{w},\mathbf{z}) + \lambda r(\mathbf{w}). \tag{1}$$

The resulting parameter vector $\hat{\mathbf{w}}_\lambda$ depends on a hyperparameter λ. This hyperparameter should be set to minimize the following equation.

$$\hat{\lambda} = \arg\min_{\lambda\in\mathbb{R}} \sum_{\mathbf{z}\in\mathcal{Z}_v} \ell(\hat{\mathbf{w}}_\lambda,\mathbf{z}) + \gamma \left\|\lambda\right\|^2. \tag{2}$$

This minimization is usually done in an ad-hoc manner via cross validation and grid search.

2 Warm up: Ridge Regression

In ridge regression, the space of data points is decomposed as $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. The loss function $\ell(\mathbf{w},(\mathbf{x},y)) = \left|\mathbf{w}^\intercal\mathbf{x} - y\right|^2$, and the regularization function $r(\mathbf{w}) = \left\|\mathbf{w}\right\|^2$. Substituting into (1) gives

$$\hat{\mathbf{w}}_\lambda^{\text{ridge}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \sum_{(\mathbf{x},y)\in\mathcal{Z}_t} \left|\mathbf{w}^\intercal\mathbf{x} - y\right|^2 + \lambda\left\|\mathbf{w}\right\|^2. \tag{3}$$

It is common to let X_t be the $n\times d$ matrix of data points in $\mathcal{X}_t$ and Y_t to be the $n\times 1$ matrix of corresponding response variables in $\mathcal{Y}_t$. Then (3) can be rewritten as

$$\hat{\mathbf{w}}_\lambda^{\text{ridge}} = \arg\min_{\mathbf{w}\in\mathcal{W}} \left\|X_t\mathbf{w} - Y_t\right\|^2 + \lambda\left\|\mathbf{w}\right\|^2. \tag{4}$$

For a fixed λ, (5) has the closed form solution

$$\hat{\mathbf{w}}_\lambda^{\text{ridge}} = \left({X_t}^\intercal X_t + \lambda I\right)^{-1} {X_t}^\intercal Y_t. \tag{5}$$

We now rewrite the equation for the optimal hyperparameter (2) as

$$\hat{\lambda}^{\text{ridge}} = \arg\min_{\lambda\in\mathbb{R}} \left\|X_v\hat{\mathbf{w}}_\lambda^{\text{ridge}} - Y_v\right\|^2 + \gamma\lambda^2 \tag{6}$$

$$= \arg\min_{\lambda\in\mathbb{R}} \left\|X_v\left({X_t}^\intercal X_t + \lambda I\right)^{-1} {X_t}^\intercal Y_t - Y_v\right\|^2 + \gamma\lambda^2. \tag{7}$$

Unlike $\hat{\mathbf{w}}_\lambda^{\text{ridge}}$, $\hat{\lambda}^{\text{ridge}}$ does not appear to have a closed form solution. The objective is neither convex nor guaranteed to have a single minima. So we turn to numerical optimization procedures.

3 Problem Setting

To solve (2) analytically, we set the derivative inside the arg min to zero and solve for λ. This gives the equation

$$0 = \frac{\mathrm{d}}{\mathrm{d}\,\lambda} \sum_{\mathbf{z} \in \mathcal{Z}_v} \ell(\hat{\mathbf{w}}_\lambda, \mathbf{z}) = \sum_{\mathbf{z} \in \mathcal{Z}_v} \frac{\partial}{\partial \hat{\mathbf{w}}_\lambda} \ell(\hat{\mathbf{w}}_\lambda, \mathbf{z}) \frac{\mathrm{d}}{\mathrm{d}\,\lambda} \hat{\mathbf{w}}_\lambda. \tag{8}$$

To solve (8), we need to calculate $\frac{\mathrm{d}}{\mathrm{d}\,\lambda}\hat{\mathbf{w}}_\lambda$. This is the derivative of the arg min function. We appeal to the following theorem.

Theorem 1 (Gould et al. (2016))**.** *Let $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}$ be a twice differentiable function. Let $g(x) = \arg\min_{y \in \mathbb{R}^n} f(x, y)$. Then,*

$$\frac{\mathrm{d}}{\mathrm{d}\,x} g(x) = \left(\nabla_y^2 f(x, y)\right)^{-1} \left(\frac{\partial}{\partial x} \nabla_y f(x, y)\right). \tag{9}$$

Applying this theorem to $\frac{\mathrm{d}}{\mathrm{d}\,\lambda}\hat{\mathbf{w}}_\lambda$ gives

$$\frac{\mathrm{d}}{\mathrm{d}\,\lambda} \hat{\mathbf{w}}_\lambda = \left(\sum_{\mathbf{z} \in \mathcal{Z}_v} \nabla_{\mathbf{w}}^2 \ell(\mathbf{w}, \mathbf{z}) + \lambda \nabla_{\mathbf{w}}^2 r(\mathbf{w})\right)^{-1} \lambda \nabla_{\mathbf{w}} r(\mathbf{w}). \tag{10}$$

Substituting (10) into (8) yields

$$0 = \left(\sum_{\mathbf{z} \in \mathcal{Z}_v} \frac{\partial}{\partial \hat{\mathbf{w}}_\lambda} \ell(\hat{\mathbf{w}}_\lambda, \mathbf{z})\right) \left(\sum_{\mathbf{z} \in \mathcal{Z}_v} \nabla_{\hat{\mathbf{w}}_\lambda}^2 \ell(\hat{\mathbf{w}}_\lambda, \mathbf{z}) + \lambda \nabla_{\hat{\mathbf{w}}_\lambda}^2 r(\hat{\mathbf{w}}_\lambda)\right)^{-1} \lambda \nabla_{\hat{\mathbf{w}}_\lambda} r(\hat{\mathbf{w}}_\lambda). \tag{11}$$

References

Stephen Gould, Basura Fernando, Anoop Cherian, Peter Anderson, Rodrigo Santa Cruz, and Edison Guo. On differentiating parameterized argmin and argmax problems with application to bi-level optimization. *arXiv preprint arXiv:1607.05447*, 2016.