Convex Techniques for Model Selection

Dustin Tran

May 8, 2014

Introduction

Regularization, and thus determining the optimum regularization parameter, occurs in many learning algorithms, e.g.,

- Tikhonov regularization
- Lasso regression
- Smoothing splines
- Regularization networks
- SVMs
- LS-SVMs

Introduction

Current methods to measure the appropriateness of a regularization parameter for given data are:

- Cross-Validation
- Generalized CV
- Mallows's C_p
- Minimum Description Length (MDL)
- AIC
- BIC

Recent interest has also been in discovering closed-form expressions of the solution path, and developing homotopy methods.

Proposal

This paper concerns a convex method for finding the optimal regularization parameter.

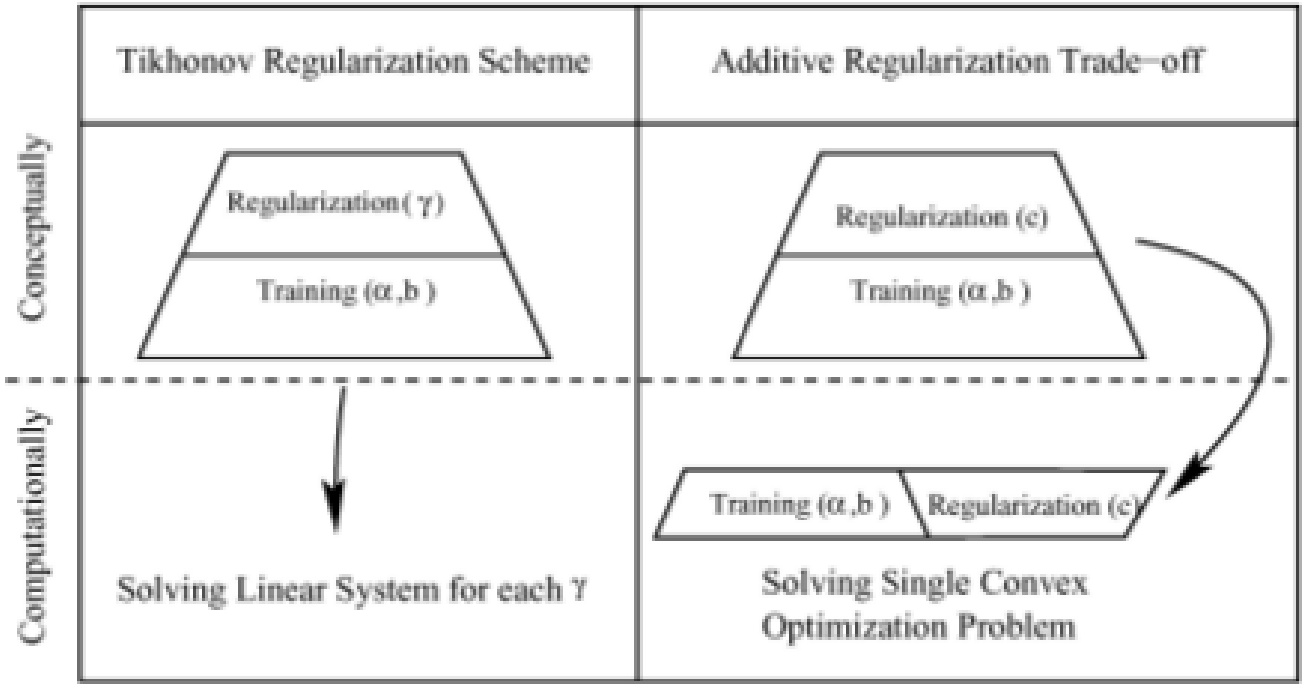

Figure 1. Comparison between the classical Tikhonov regularization scheme and the additive regularization trade-off scheme: Conceptually, training and validation levels are different in both schemes. Computationally, fusion of the training and validation levels results in a single constrained optimization problem. In the case of the additive regularization trade-off this problem becomes a convex problem after fusion.

Advantages

- **Automation**: Practical users of machine learning tools may not be not interested in tuning the parameter manually. This brings us closer to fully automated algorithms
- **Convexity**: It is (usually) much easier to examine worst case behavior for convex sets, than attempting to characterize all possible local minima
- **Extensibility**: This method applied to ridge regression here naturally extends to more complex model selection problems (e.g. backward selection), which often suffer from local minima and the lack of a formal framework
- **Performance**: The algorithmic approach of first finding the solution of the convex relaxation and then projecting it to the original non-convex solution path is—in particular scenarios—more efficient than general purpose optimization routines

Ridge Solution Set

Let $v \in \mathbb{R}^n$ and M be a $n \times n$ positive semi-definite matrix. For a fixed $\gamma \in (0, \infty)$, Tikhonov regularization schemes of linear operators lead to the solution $\hat{u} \in \mathbb{R}^n$, where the estimator $\hat{u}$ solves

$$(M + \gamma I_n)u = v \tag{1}$$

Definition

For a fixed value $\gamma_0 > 0$, we define the ridge solution set S as the set of all solutions $\hat{u}$ corresponding to a value $\gamma \in (\gamma_0, \infty)$. That is,

$$S(\gamma, u|\gamma_0, M, v) := \{\gamma \in (\gamma_0, \infty),\ u \in \mathbb{R}^n | (M + \gamma I_n)u = v\}. \tag{2}$$

Let $M = U\Sigma U^T$ denote the SVD of M, i.e., U is orthogonal and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_n)$ contains all the ordered positive eigenvalues with $\sigma_1 \geq \cdots \geq \sigma_n$.

Convex Relaxation to Ridge

Proposition

Set $\sigma_i' := \sigma_i + \gamma_0$ *for all* $i = 1, \ldots, n$. *Then the polytope* $\mathcal{R}$ *parametrized by* $\Lambda = \{\lambda_1, \ldots, \lambda_n\}$ *as follows*

$$\mathcal{R}(\Lambda, u|\gamma_0, M, v) = \begin{cases} U_i^T u = \lambda_i U_i^T v, & \text{for all } i = 1, \ldots, n \\ 0 < \lambda_i < \frac{1}{\sigma_i'}, & \text{for all } i = 1, \ldots, n \\ \frac{\sigma_k'}{\sigma_i'} \lambda_k \leq \lambda_i \leq \lambda_k, & \text{for all } \sigma_i' > \sigma_k' \\ \lambda_i = \lambda_k, & \text{for all } \sigma_i' = \sigma_k' \end{cases} \tag{3}$$

is convex, and moreover, forms a convex hull to S.

The proof is somewhat tedious and long, and of which we omit for this poster session.

Original Method

Let $\mathcal{D} = \{(x_{i1}, \ldots, x_{ip}, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ be a given data set. The ridge regression estimator $f(x) = x^T\beta$ with $\beta \in \mathbb{R}^p$ minimizes the regularized loss function

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell(y_i - x_i^T\beta) + \frac{\gamma}{2}\|\beta\|_2^2 \tag{4}$$

where $\ell : \mathbb{R} \to \mathbb{R}_{>0}$ is some loss function.

Proposition (Ridge Regression)

Set $\ell(z) := z^2$ and fix $\gamma > 0$. For $\beta \in \mathbb{R}^p$ to be the unique global minimizer of (4), it is necessary and sufficient that β satisfies

$$KKT(\beta|\gamma, \mathcal{D}) \ : \ (X^TX + \gamma I_p)\beta = X^Ty, \tag{5}$$

where X is the $n \times p$ design matrix and $y \in \mathbb{R}^n$ is the response vector formulated from $\mathcal{D}$, and I_p is the $p \times p$ identity matrix.

Original Method

Note that we use the notation of KKT in order to hint to the extension to other learning machines (as SVMs), which reduce to solving a similar convex optimization problem but with inequality constraints.

Let $\mathcal{D}^v = \{(x_{i1}^v, \ldots, x_{ip}^v, y_i^v)\}_{i=1}^{n_v} \subset \mathbb{R}^p \times \mathbb{R}$ be a validation data set. The optimization problem of finding the optimal regularization parameter $\hat{\beta}$ with respect to a validation performance criterion $\hat{\gamma}$ can then be written as

$$(\hat{\beta}, \hat{\gamma}) = \underset{\beta, \gamma > 0}{\arg\min} \sum_{j=1}^{n_v} \ell(y_j^v - (x_j^v)^T \beta) \text{ s.t. } KKT(\beta|\gamma, \mathcal{D}) = S(\gamma, \beta|\mathcal{D}) \quad (6)$$

That is, we find the least squared error among all β's in the solution set S, or equivalently all β's satisfying the KKT conditions.

Convex Method

For the convex approach, we simply replace the non-convex solution set $S(\gamma, \beta|\mathcal{D})$ with its convex relaxation $\mathcal{R}(\Lambda, \beta|\mathcal{D})$. Then one obtains the convex optimization problem

$$(\hat{\beta}, \hat{\Lambda}) = \arg\min_{\beta,\Lambda} \sum_{j=1}^{n_v} \ell(y_j^v - (x_j^v)^T \beta) \text{ s.t. } R(\Lambda, \beta|\mathcal{D}) \tag{7}$$

This has the immediate advantage of simultaneously training and validating (1 step); the original method requires finding a grid of points $(\hat{\beta}, \hat{\gamma})$ in S and then minimizing among those (2 steps).

For example, (7) can be solved with a QP solver when $\ell(z) = z^2$ as before, or with a LP solver when $\ell(z) = |z|$ (the latter of which may be preferred from a robustness or computational point of view).

Convex Method

Corollary

The convex relaxation constitutes the solution path for the modified ridge regression problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} \ell(y_i - x_i^T \beta) + \frac{1}{2}\beta^T (U \Gamma U^T)\beta \tag{8}$$

where $\Gamma = \mathrm{diag}(\gamma_1, \ldots, \gamma_p)$ *and* γ_k *satisfies the constraint* $\gamma_k = \frac{1}{\lambda_k} - \sigma_k$ *for all* $k = 1, \ldots, p$*, and the following inequalities hold by translating (3):*

$$\begin{cases} \gamma_k > 0, & \text{for all } k = 1, \ldots, p \\ \frac{\sigma_\ell}{\sigma_k}(\sigma_k + \gamma_k) \geq \sigma_\ell + \gamma_\ell > \sigma_k + \gamma_k, & \text{for all } \sigma_\ell > \sigma_k \\ \gamma_\ell = \gamma_k, & \text{for all } \sigma_\ell = \sigma_k \end{cases} \tag{9}$$

Extension to K-fold Cross-Validation

Let $\mathcal{D}_{(k)}$ and $\mathcal{D}^v_{(k)}$ denote the set of training and validation data respectively, corresponding to the k^{th} fold for $k = 1, \ldots, K$: that is, they satisfy

$$\bigcup_k \mathcal{D}^v_{(k)} = \mathcal{D}, \qquad \bigcap_k \mathcal{D}_{(k)} = \bigcap_k \mathcal{D}^v_{(k)} = \emptyset \tag{10}$$

Let $n_{(k)} = |\mathcal{D}_{(k)}|$. Then in order to tune the parameter γ according to K-fold CV, we have the optimization problems

$$(\hat{\beta}_{(k)}, \hat{\gamma}) = \underset{\beta_{(k)}, \gamma > 0}{\arg\min} \sum_{k=1}^{K} \frac{1}{n - n_{(k)}} \sum_{(x_j, y_j) \in \mathcal{D}^v_{(k)}} \ell(y_j - x_j^T \beta_{(k)})$$

$$\text{s.t. } KKT(\beta_{(k)} | \gamma, \mathcal{D}_{(k)}) = S(\gamma, \beta_{(k)} | \mathcal{D}_{(k)}) \tag{11}$$

for all $k = 1, \ldots, K$.

Extension to K-fold Cross-Validation

Then we need only relax the KKT conditions independently for each k. The convex optimization problem according to a K-fold CV is

$$(\hat{\beta}_{(k)}, \hat{\Lambda}) = \underset{\beta_{(k)}, \Lambda^k}{\arg\min} \sum_{k=1}^{K} \frac{1}{n - n_{(k)}} \sum_{(x_j, y_j) \in \mathcal{D}^v_{(k)}} \ell(y_j - x_j^T \beta_{(k)})$$
$$\text{s.t. } \mathcal{R}(\Lambda, \beta_{(k)} | \mathcal{D}_{(k)}) \tag{12}$$

for all $k = 1, \ldots, K$, each of which is solved as before.

Then just as in typical K-fold CV, we take the average of the folds $\hat{\beta}_{avg} = \frac{1}{K} \sum_{k=1}^{K} \hat{\beta}_{(k)}$ as the final model.

Performance

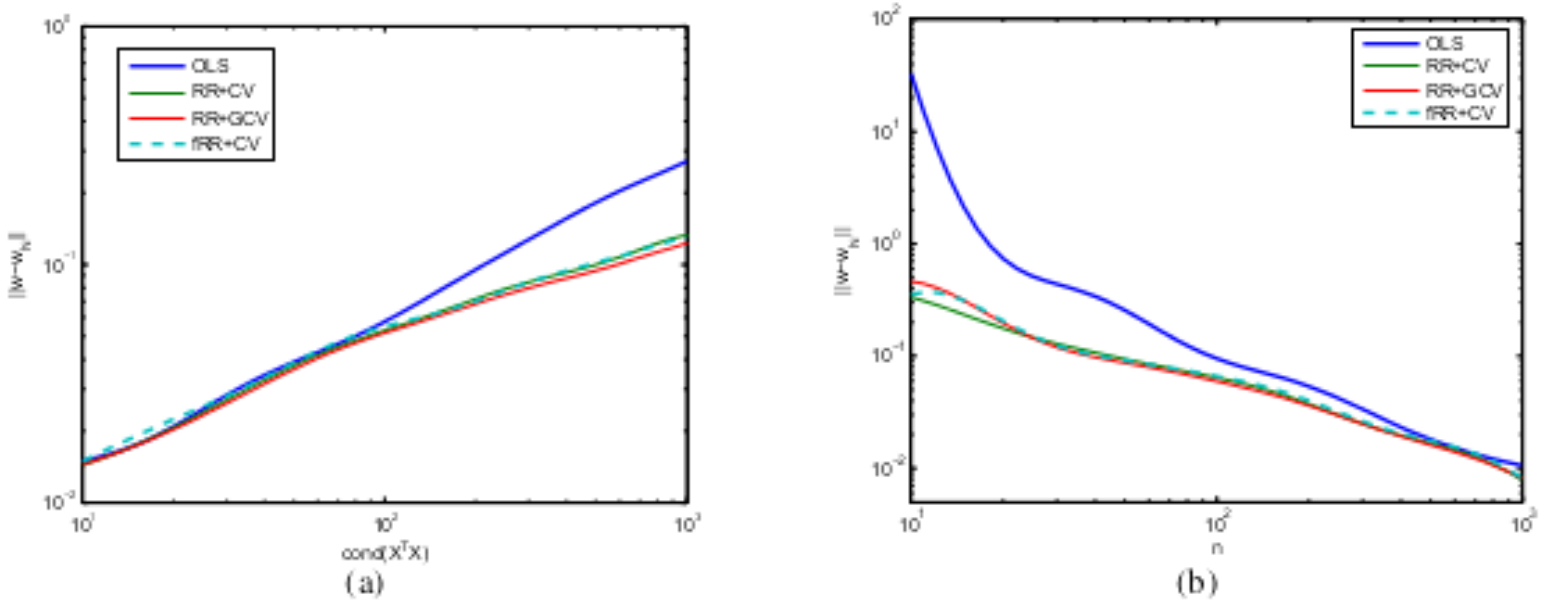

Figure 1: *Results of a comparison between OLS and RR with $D = 10$, tuned by CV (steepest descent), by GCV (using steepest descent) and the proposed method fusing training and tuning the ridge together in one convex optimization algorithm. Panel* **(a)** *shows the evolution when ranging the condition number with $n = 50$ fixed. Panel* **(b)** *displays the evolution of the performance when the number of examples ranges and $\Gamma(X^T X) = 1e^3$ is fixed. In both cases the proposed convex relaxation is performing similar as steepest descent based counterparts, while it significantly outperforms OLS in the case of low n or a high enough condition number $\Gamma(X^T X)$.*

Performance