

CONVEX TECHNIQUES FOR MODEL SELECTION

DUSTIN TRAN

ABSTRACT. We develop a robust convex algorithm to select the regularization parameter in model selection. In practice this would be automated in order to save practitioners time from having to tune it manually. In particular, we implement and test the convex method for K -fold cross validation on ridge regression, although the same concept extends to more complex models. We then compare its performance with standard methods.

1. INTRODUCTION

Regularization has been studied extensively since the theory of ill-posed problems [6]. By adding a penalty associated to a choice of regularization parameter, one penalizes overfitted models and can achieve good generalized performance with the training data. This has been a crucial feature in current modeling frameworks: for example, Tikhonov regularization [10], Lasso regression [5], smoothing splines [8], regularization networks [1], SVMs [11], and LS-SVMs [9]. SVMs in particular are characterized by dual optimization reformulations, and their solutions follow from convex programs. Standard SVMs reduce to solving quadratic problems and LS-SVMs reduce to solving a set of linear equations.

Many general purpose methods to measure the appropriateness of a regularization parameter for given data exist: cross-validation (CV), generalized CV, Mallows's C_p , Minimum Description Length (MDL), Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC). Recent interest has also been in discovering closed-form expressions of the solution path [3], and developing homotopy methods [7].

Like SVMs, this paper takes the perspective of convex optimization—following [2]—in order to tune the regularization parameter for optimal model selection. Classical Tikhonov regularization schemes require two steps: 1. (training) choosing a grid of fixed parameter values, find the solution for each constant regularization parameter; 2. (validating) optimize over the regularization constants and choose the model according to a model selection criterion. The approach in this paper reformulates the problem as one for constrained optimization. This allows one to compute both steps simultaneously: minimize the validation measure subject to the training equations as constraints.

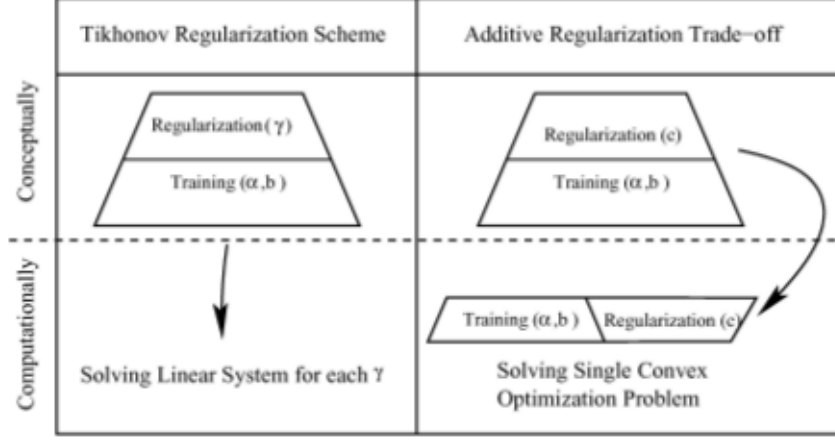


Figure 1. Comparison between the classical Tikhonov regularization scheme and the additive regularization trade-off scheme: Conceptually, training and validation levels are different in both schemes. Computationally, fusion of the training and validation levels results in a single constrained optimization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

This approach offers noticeable advantages over the ones above, of which we outline a few here:

- **Automation:** Practical users of machine learning tools may not be 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, rather than attempting to characterize all possible local minima.
- **Performance:** The algorithmic approach of training and validating simultaneously is occasionally more efficient than general purpose optimization routines.

For this write-up, we focus only on ridge regression, although the same approach can be applied to more complex model selection problems (which may benefit more as they suffer more often from local minima).

2. BACKGROUND

2.1. Ridge solution set. We introduce the standard approach to ridge regression and prove key properties for a convex reformulation. For a more rigorous introduction, see [4]. Let $y \in \mathbb{R}^n$ and M be a $n \times n$ positive semi-definite matrix. For a fixed $\gamma \in (0, \infty)$, recall that Tikhonov regularization schemes of linear operators lead to the solution $\hat{\beta} \in \mathbb{R}^n$, where the estimator $\hat{\beta}$ solves

$$(2.1) \quad (M + \gamma I_n)\beta = y$$

Definition 2.1. For a fixed value $\gamma_0 > 0$, we define the ridge solution set S as the set of all solutions $\hat{\beta}$ corresponding to a value $\gamma \in (\gamma_0, \infty)$. That is,

$$(2.2) \quad S(\gamma, \beta | \gamma_0, M, y) := \{\gamma \in (\gamma_0, \infty), \beta \in \mathbb{R}^n | (M + \gamma I_n)\beta = y\}$$

The value γ_0 can be thought of as the minimal regularization parameter allowed in the solution set; this would speed up computation should the user already know a lower bound on their optimal choice of γ .

Let $M = U\Sigma U^T$ denote the reduced singular value decomposition (SVD) of M , i.e., U is orthogonal and $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_n)$ contains all the ordered positive eigenvalues with $\sigma_1 \geq \dots \geq \sigma_n$.

Lemma 2.2. The solution function $h_{M,y}(\gamma) = (M + \gamma I_n)^{-1}y$ is Lipschitz continuous.

Proof. For two values γ_1, γ_2 such that $\gamma_0 \leq \gamma_1 \leq \gamma_2 < \infty$,

(2.3)

$$\|(M + \gamma_1 I_n)^{-1}y - (M + \gamma_2 I_n)^{-1}y\|_2 = \left\| U \left((\Sigma + \gamma_1 I_n)^{-1} - (\Sigma + \gamma_2 I_n)^{-1} \right) U^T y \right\|_2$$

$$(2.4) \quad \leq \max_i \left| \frac{1}{\sigma_i + \gamma_1} - \frac{1}{\sigma_i + \gamma_2} \right| \|y\|_2$$

$$(2.5) \quad \leq \left| \frac{1}{\sigma_n + \gamma_1} - \frac{1}{\sigma_n + \gamma_2} \right| \|y\|_2$$

Consider the function

$$(2.6) \quad g(x) = \frac{1}{\sigma_n + \gamma_0 + x}$$

Then by the mean value theorem, there exists a value $c \in [\lambda_1 - \lambda_0, \lambda_2 - \lambda_0]$ such that

$$(2.7) \quad \left| \frac{1}{\sigma_n + \lambda_1} - \frac{1}{\sigma_n + \lambda_2} \right| \|y\|_2 \leq |g'(c)| |\lambda_1 - \lambda_2| \|y\|_2$$

$$\leq \frac{\|y\|_2}{(\sigma_n^2 + \lambda_0)^2} |\lambda_1 - \lambda_2| \quad \square$$

2.2. Convex relaxation. We now examine the the convex hull of the solution set $S(\gamma, \beta | \gamma_0, M, y)$. This will allow us to search efficiently through a convex set of hypotheses as in Section 3. We decompose the solution function into a sum of low rank matrices:

$$(2.8) \quad h_{M,y}(\gamma) = (M + \gamma I_n)^{-1}y = U(\Sigma + \gamma I_n)^{-1}U^T y = \sum_{i=1}^n \frac{1}{\sigma_i + \gamma} U_i U_i^T y$$

Then define $\lambda_i := 1/(\sigma_i + \gamma)$ for all $i = 1, \dots, n$. The following proposition provides linear constraints on the set of λ_i 's following from this reparameterization.

Proposition 2.3. *Let $\sigma'_i := \sigma_i + \gamma_0$ for $i = 1, \dots, n$. Then the polytope \mathcal{R} parametrized by $\Lambda = \{\lambda_1, \dots, \lambda_n\}$ as follows*

$$(2.9) \quad \mathcal{R}(\Lambda, \beta|\gamma_0, M, y) = \begin{cases} U_i^T \beta = \lambda_i U_i^T y, & \text{for all } i = 1, \dots, n \\ 0 < \lambda_i \leq \frac{1}{\sigma'_i}, & \text{for all } i = 1, \dots, n \\ \frac{\sigma'_{i+1}}{\sigma'_i} \lambda_{i+1} \leq \lambda_i \leq \lambda_{i+1}, & \text{for all } \sigma'_i \geq \sigma'_{i+1}, i = 1, \dots, n-1 \end{cases}$$

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

Proof. It is easy to verify the first two constraints by looking at the function g defined previously, which is strictly increasing. Set $\gamma' := \gamma - \gamma_0 > 0$, and $\lambda_i = 1/(\sigma'_i + \gamma')$ as before. Then for all $\sigma'_i \geq \sigma'_{i+1}$,

$$(2.10) \quad \lambda_i = \left(\frac{1}{\sigma'_{i+1} + \gamma'} \right) / \left(\frac{\sigma'_{i+1} + \gamma' + \sigma'_i - \sigma'_{i+1}}{\sigma'_{i+1} + \gamma'} \right)$$

$$(2.11) \quad = \frac{\lambda_{i+1}}{1 + \lambda_{i+1}(\sigma'_i - \sigma'_{i+1})}$$

$$(2.12) \quad \geq \frac{\lambda_{i+1}}{1 + \frac{1}{\sigma'_{i+1}}(\sigma'_i - \sigma'_{i+1})} = \frac{\sigma'_{i+1}}{\sigma'_i} \lambda_{i+1}$$

Hence the third inequality is also true. Moreover, the set $\mathcal{R}(\Lambda, \beta|\gamma_0, M, y)$ is characterized entirely by equalities and inequalities which are linear in the unknown λ_i 's. So it forms a polytope. Then by the above result together with the convex property of polytopes, it follows that $\mathcal{R}(\Lambda, \beta|\gamma_0, M, y)$ is a convex relaxation to the set $S(\gamma, \beta|\gamma_0, M, y)$. \square

We now find the relationship between solutions in $S(\gamma, \beta|\gamma_0, M, y)$ and solutions in its convex relaxation $\mathcal{R}(\Lambda, \beta|\gamma_0, M, y)$.

Proposition 2.4. *The maximal distance between a solution in $\mathcal{R}(\Lambda, \beta|\gamma_0, M, y)$ and its closest counterpart in $S(\gamma, \beta|\gamma_0, M, y)$ is bounded by the maximum range of the inverse eigenvalue spectrum:*

$$(2.13) \quad \min_{\gamma} \|U \Sigma U^T y - (M + \gamma I_n)^{-1} y\|_2 \leq \|y\|_2 \max_{i > j} \left| \frac{1}{\sigma'_i} - \frac{1}{\sigma'_j} \right|$$

Proof. Following similar steps to prove Lemma 2.2, one can show that the maximal difference between a solution β_Γ for a given Γ and its corresponding closest $\beta_{\hat{\gamma}}$ is

$$(2.14) \quad \min_{\gamma} \|U \Lambda U^T y - (M + \hat{\gamma} I_n)^{-1} y\|_2 \leq \min_{\gamma} \|y\|_2 \max_{i=1, \dots, n} \left| \lambda_i - \frac{1}{\sigma'_i + \gamma} \right|$$

For any $\lambda_i \neq \lambda_j$, the value

$$(2.15) \quad \min_{\gamma} \max \left\{ \left| \lambda_i - \frac{1}{\sigma'_i - \gamma'} \right|, \left| \lambda_j - \frac{1}{\sigma'_j - \gamma'} \right| \right\}$$

is bounded by the worst case scenario that the solution γ passes through λ_i or through λ_j . Then

$$(2.16) \quad \min_{\gamma} \max_i \left| \lambda_i - \frac{1}{\sigma'_i + \gamma} \right| \leq \max_{i \neq j} \left| \lambda_i - \frac{1}{\sigma'_i + \gamma_j} \right|$$

$$(2.17) \quad < \max_{i \neq j} \lambda_i \left| 1 - \frac{\sigma'_i}{\sigma'_j} \right|$$

$$(2.18) \quad < \max_{i > j} \left| \frac{1}{\sigma'_i} - \frac{1}{\sigma'_j} \right|$$

where γ_j satisfies $\lambda_j = 1/(\sigma'_j + \gamma_j)$. That is, $\gamma_j = 1/\lambda_j - \sigma'_j$ which is greater than zero by construction. Hence for all $\Gamma \in \mathbb{R}^n$, there exists $\hat{\gamma}$ such that

$$(2.19) \quad \|\beta_{\Gamma} - \beta_{\hat{\gamma}}\|_2 \leq \|y\|_2 \max_{i > j} \left| \frac{1}{\sigma'_i} - \frac{1}{\sigma'_j} \right| \quad \square$$

2.3. Reformulation of ridge regression. Let $\mathcal{D} = \{(x_{i1}, \dots, x_{ip}, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ 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

$$(2.20) \quad \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \ell(y_i - x_i^T \beta) + \frac{\gamma}{2} \|\beta\|_2^2$$

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

Proposition 2.5 (KKT conditions). *Set $\ell(z) := z^2$ and fix $\gamma > 0$. For $\beta \in \mathbb{R}^p$ to be the unique global minimizer of (2.20), it is necessary and sufficient that β satisfies*

$$(2.21) \quad KKT(\beta|\gamma, \mathcal{D}) : (X^T X + \gamma I_p) \beta = X^T y,$$

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

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

Let $\mathcal{D}^v = \{(x_{i1}^v, \dots, x_{ip}^v, y_i^v)\}_{i=1}^{n_v} \subset \mathbb{R}^p \times \mathbb{R}$ 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

$$(2.22) \quad (\hat{\beta}, \hat{\gamma}) = \arg \min_{\beta, \gamma > 0} \sum_{j=1}^{n_v} \ell(y_j^v - (x_j^v)^T \beta) \text{ s.t. } KKT(\beta|\gamma, \mathcal{D}) = S(\gamma, \beta|\mathcal{D})$$

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.

3. CONVEX PROGRAM

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

$$(3.1) \quad (\hat{\beta}, \hat{\Lambda}) = \arg \min_{\beta, \Lambda} \sum_{j=1}^{n_v} \ell(y_j^v - (x_j^v)^T \beta) \text{ s.t. } R(\Lambda, \beta|\mathcal{D})$$

This has the immediate advantage of simultaneously training and validating (1 step); in comparison the original method requires finding a grid of points $(\hat{\beta}, \hat{\gamma})$ in S and then minimizing among those (2 steps). Furthermore, the convex hull is defined by $\mathcal{O}(n)$ equality/inequality constraints, whose complexity is not any higher than the original problem.

For example, (3.1) 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 for sparsenesses or feature selection).

Corollary 3.1. *The convex relaxation constitutes the solution path for the modified ridge regression problem*

$$(3.2) \quad \hat{\beta} = \arg \min_{\beta} \sum_{i=1}^n \ell(y_i - x_i^T \beta) + \frac{1}{2} \beta^T (U \Gamma U^T) \beta$$

where $\Gamma = \text{diag}(\gamma_1, \dots, \gamma_p)$ and $\gamma_i = \frac{1}{\lambda_k} - \sigma_k$ for all $i = 1, \dots, p$, and the following inequalities hold by translating (2.9):

$$(3.3) \quad \begin{cases} \gamma_i > 0, & \text{for all } i = 1, \dots, p \\ \frac{\sigma_{i+1}}{\sigma_i}(\sigma_i + \gamma_i) \geq \sigma_{i+1} + \gamma_{i+1} > \sigma_i + \gamma_i, & \text{for all } \sigma_{i+1} \geq \sigma_i, i = 1, \dots, p-1 \end{cases}$$

3.1. Generalization to K -fold cross-validation. The above applies to a single training and validation set, and we now extend it to K -fold CV in general. Let $\mathcal{D}_{(k)}$ and $\mathcal{D}_{(k)}^v$ denote the set of training and validation data respectively, corresponding to the k^{th} fold for $k = 1, \dots, K$: that is, they satisfy

$$(3.4) \quad \bigcup_k \mathcal{D}_{(k)}^v = \mathcal{D}, \quad \bigcap_k \mathcal{D}_{(k)} = \bigcap_k \mathcal{D}_{(k)}^v = \emptyset$$

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

$$(3.5) \quad \begin{aligned} (\hat{\beta}_{(k)}, \hat{\gamma}) &= \arg \min_{\beta_{(k)}, \gamma > 0} \sum_{k=1}^K \frac{1}{n - n_{(k)}} \sum_{(x_j, y_j) \in \mathcal{D}_{(k)}^v} \ell(y_j - x_j^T \beta_{(k)}) \\ &\text{s.t. } KKT(\beta_{(k)}|\gamma, \mathcal{D}_{(k)}) = S(\gamma, \beta_{(k)}|\mathcal{D}_{(k)}) \end{aligned}$$

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

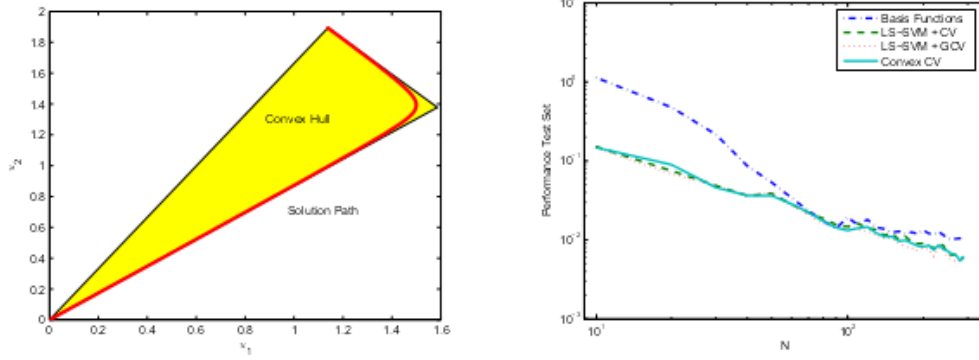
Then we need only relax the KKT conditions independently for each k . The convex optimization problem according to a K -fold CV is

$$(3.6) \quad \begin{aligned} (\hat{\beta}_{(k)}, \hat{\Lambda}) = \arg \min_{\beta_{(k)}, \Lambda^k} & \sum_{k=1}^K \frac{1}{n - n_{(k)}} \sum_{(x_j, y_j) \in \mathcal{D}_{(k)}^v} \ell(y_j - x_j^T \beta_{(k)}) \\ \text{s.t. } & \mathcal{R}(\Lambda, \beta_{(k)} | \mathcal{D}_{(k)}) \end{aligned}$$

for all $k = 1, \dots, K$, each of which is solved as before, and so with $\mathcal{O}(Kn)$ constraints. 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.

4. PERFORMANCE

We show two simulation studies as a benchmark in order to compare it to current methods. The first figure below provides intuition behind the solution paths: the curve $S(\lambda, \beta | \gamma_0, M, y)$ and its convex relaxation $\mathcal{R}(\Lambda, \beta, | \gamma_0, M, y)$. In the second figure, we simulate data with the function $y = \text{sinc}(x) + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$ are sampled i.i.d., $n = 50$ observations, and $p = 5$.



The figure compares the performance of the method with one which uses basis functions, another with CV, and another with generalized CV (GCV). CV and GCV are implemented with standard gradient descent, and the convex algorithm outlined here uses the interior point method. This toy example demonstrates that the relaxation does not result in an increase in true error.

We also conduct a Monte Carlo simulation: every iteration constructs a simulated model for a given value of n defined as

$$(4.1) \quad f_p(x) = w_1 x^1 + \dots + w_p x^p$$

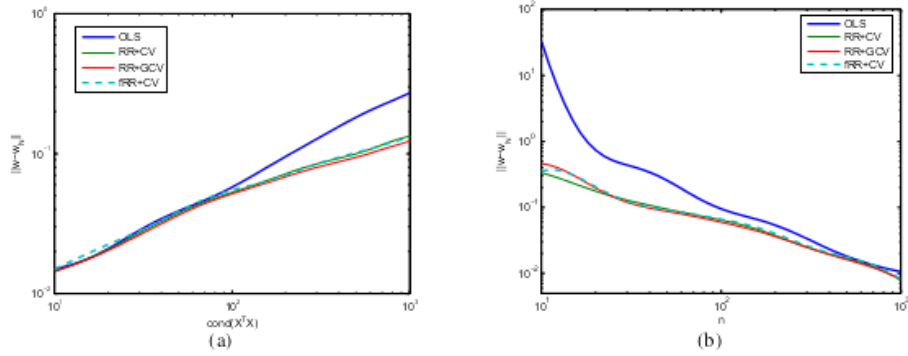
for random values of $w = (w_1, \dots, w_p) \in \mathbb{R}^p$ where $p = 10$, $w \sim \mathcal{N}(0, C)$, and C is a $p \times p$ covariance matrix with $\|C\|_2 = 1$. A data set of size n is constructed such

that $\mathcal{D} = \{(x_{i1}, \dots, x_{ip}, y_i)\}_{i=1}^n$ and

$$(4.2) \quad y_i = f_m(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, 1)$$

is sampled i.i.d. for all $i = 1, \dots, n$.

We compare three methods for tuning the parameter with respect to the ordinary least squares (OLS) estimate: 10-fold CV with gradient descent (**RR+CV**), generalized CV with gradient descent (**RR+GCV**), and 10-fold CV criterion which applies the convex method as in (18) (**fRR+CV**). We run it for 20,000 iterations.



The figure on the left compares the true error as the condition number grows; the figure on the right compares the true error as the number of observations n increases.

As we can see, the method is comparable to both CV and GCV for variable changes in the data set. We expect that the OLS worsens drastically over ill-conditioned matrices, and our method compensates for that via the optimal tuning parameter which is roughly the same as CV's.

5. CONCLUSION

Viewing the model selection problem as one in constrained optimization gives rise to a natural approach to tuning the regularization parameter. According to the simulation results, the convex program provides comparable performance to popular methods. Moreover, global optimality is guaranteed and efficient convex algorithms can be employed as suited to the modeling problem at hand. This would especially outperform general purpose techniques when there is a high number of local optima, or if finding each individual solution for fixed regularization parameter is more costly than optimizing it simultaneously with validation as we do here.

Further extensions to this framework can be used as a generic convex hull method [12], and it can also be applied for constructing stable kernel machines, feature selection, and other possibilities related to simultaneous training and validating.

REFERENCES

- [1] M. Bertero, T. Poggio, and T. Torre. Ill-posed problems in early vision. *Proceedings of the IEEE*, 76(8):869–889, 1988.
- [2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.
- [4] G. Golub, M. Heath, G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21:215–223, 1979.
- [5] A. Hoerl and R. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):5582, 1970.
- [6] V. Ivanov. *The Theory of Approximate Methods and Their Application to the Numerical Solution of Singular Integral Equations*. Nordhoff International Publishing, 1976.
- [7] M.R. Osborne, B. Presnell, and B.A. Turlach. On the LASSO and its dual. *Journal of Computational & Graphical Statistics*, 2000.
- [8] L. Schumacher. *Spline Functions: Basic Theory*. New York: Wiley, 1981.
- [9] J. Suykens, and J. Vandewalle. Least squares support vector machine classifiers. *Neural Processing Letters*, 9:293–300, 1999.
- [10] A. Tikhonov and V. Arsenin. *Solutions of Ill-posed problems*. Hallsted Press, 1977.
- [11] V. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, 1998.
- [12] K. Meers, E. Ceulmans, and T. Wilderjans. CHull: A generic convex-hull-based model selection method. *Behavior research methods*, 45(1):1–15, 1977.