CS754 Assignment-4

Saksham Rathi, Ekansh Ravi Shankar, Kshitij Vaidya

Declaration: The work submitted is our own, and we have adhered to the principles of academic honesty while completing and submitting this work. We have not referred to any unauthorized sources, and we have not used generative AI tools for the work submitted here.

Question 1

Solution

1 Experimental Results

We generated a sensing matrix over a Bernoulli Distribution and a sparse signal over a Uniform Distribution. We then reconstruct the sparse signal using the CVZ package in MATLAB. The results are shown in figures 1 and 2. From the plots for VE and RMSE, we can see that they follow similar shapes and trends for different values of sparsity. This means that our hypothesis that the validation error is a good proxy for the RMSE is correct. From the plots, the optimal value of λ is in the range of $5 - 10$.

1. **Shape of the Plots** : Both the plots have a U-shaped curve. For very small values of λ, the model is underfitting, leading to high validation error and RMSE. As λ increases, the model complexity decreases, leading to lower validation error and RMSE. However, after a certain point, increasing λ leads to overfitting, causing both metrics to increase again.
2. **Effect of Sparsity** : We expect that as the sparsity increases, the values of the validation error and RMSE to be higher for sparser signals, as they are harder to reconstruct accurately. This trend is not quite exactly followed but we can attribute this to the randomness in the generation of the sensing matrix. But the overall trend and the choice of λ is still valid.

2 Effect of $\mathcal{V} = \mathcal{R}$ in Compressive Sensing

When the validation set $\mathcal{V}$ is equal to the training set $\mathcal{R}$, the theoretical guarantees and practical utility of the cross-validation in Compressive Sensing collapse.

2.1 Breakdown of Statistical Assumptions

The paper's analysis assumes that the training and validation sets are independent. When $\mathcal{V} = \mathcal{R}$ and the sets become coincident:

1. **Loss of Independence** : The Cross Validation residual $\epsilon_{\text{cv}} = \|\mathbf{y}_{cv} - \mathbf{A}_{cv}\hat{\mathbf{x}}\|_2^2$ uses the same measurements as the ($\mathbf{y}_{cv} = \mathbf{y}$, $\mathbf{A}_{cv} = \mathbf{A}$) reconstruction. This violates the statistical independence required for unbiased error estimation.
2. **Overfitting Risk** : The reconstruction algorithm minimizes $\|\mathbf{y} - \mathbf{A}\hat{\mathbf{x}}\|_2^2$. If $V = R$, ϵ_{cv} reflects a zero training error not a generalisation error leading to overfitting. The CV residual becomes a biased optimistic estimate of the true error $\varepsilon_{\mathbf{x}}$

2.2 Theoretical Guarantees Fail

The results from Lemma 1 and Theorem 1 rely on the independence of the training and validation sets. When $\mathcal{V} = \mathcal{R}$, the guarantees provided by these lemmas and theorems become invalid. The bounds on the expected error and the convergence rates are no longer applicable, as they depend on the assumption of independent samples.

1. **Lemma 1 : CV Residual Distribution** : The distribution $\epsilon_{\text{cv}} \sim \mathcal{N}\left(\mu, \sigma^2\right)$ assumes independence of $\mathcal{R}$ and $\mathcal{V}$. When $\mathcal{V} = \mathcal{R}$, the distribution of the CV residual becomes degenerate, leading to unreliable estimates of the mean and variance because the Central Limit Theorem approximation is invalidated.
2. **Theorem 1 : Expected Error Bound** : The error bound

   $$h(\lambda, \pm)\epsilon_{\text{cv}} - \sigma_{\mathbf{n}}^2 \leq \varepsilon_{\mathbf{x}} \leq h(\lambda, -)\epsilon_{\text{cv}} - \sigma_{\mathbf{n}}^2$$

   depends on ϵ_{cv} being unbiased. When $V = R$, ϵ_{cv} is downward-biased, rendering the bounds meaningless.

3 Justification of CV Residual as Proxy for MSE

The proxying ability of cross-validation residuals to estimate the actual mean squared error is addressed in Lemma 1 and Theorem 1 of the paper.

3.1 Lemma 1 : CV Residual Distribution

The CV residual $\epsilon_{\mathbf{cv}}$ is shown to follow a normal distribution parameterised by the true recovery error (MSE) $\varepsilon_{\mathbf{x}}$ and the noise variance $\sigma_{\mathbf{n}}^2$. The lemma states that:

$$\epsilon_{\text{cv}} \sim \mathcal{N}\left(\mu, \sigma^2\right),$$

where:

1. $\mu = \frac{m_{\text{cv}}}{m}\left(\varepsilon_{\mathbf{x}} + \sigma_{\mathbf{n}}^2\right)$ is the mean,
2. $\sigma^2 = \frac{2m_{\text{cv}}}{m^2}\left(\varepsilon_{\mathbf{x}} + \sigma_{\mathbf{n}}^2\right)^2$ is the variance,
3. $\varepsilon_{\mathbf{x}} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2$ is the true MSE between the original signal $\mathbf{x}$ and its estimate $\hat{\mathbf{x}}$,
4. $\sigma_{\mathbf{n}}^2$ is the noise variance,
5. m_{cv} and m are the number of CV and reconstruction measurements, respectively.

This implies that the observable CV residual is directly related to the unobservable MSE. The mean of ϵ_{cv} scales linearly with $\varepsilon_{\mathbf{x}}$, making it a statistically valid proxy.

3.2 Theorem 1 : Probabilistic Bound on Recovery Error

The theorem quantifies how tightly ϵ_{cv} bounds $\varepsilon_{\mathbf{x}}$ with high probability:

$$h(\lambda, +)\epsilon_{\text{cv}} - \sigma_{\mathbf{n}}^2 \leq \varepsilon_{\mathbf{x}} \leq h(\lambda, -)\epsilon_{\text{cv}} - \sigma_{\mathbf{n}}^2,$$

where:

1. $h(\lambda, \pm) = \dfrac{m}{m_{\text{cv}}} \dfrac{1}{1 \pm \lambda \sqrt{\dfrac{2}{m_{\text{cv}}}}}$ defines the scaling factors,
2. λ controls the confidence level via the error function $\operatorname{erf}(\lambda/\sqrt{2})$,
3. The bounds hold with probability $\operatorname{erf}(\lambda/\sqrt{2})$, e.g., $\lambda = 3$ corresponds to 99.7% confidence.

3.3 Conclusions

1. **Proxy Relationship** : The inequality directly uses ϵ_{cv} to bound $\varepsilon_{\mathbf{x}}$, even though $\mathbf{x}$ is unknown. This enables us to estimate recovery quality using only the CV residual.
2. **Tightness of Bounds:** The interval width scales as $\mathcal{O}(1/m_{\text{cv}}^{2/3})$, meaning more CV measurements (m_{cv}) tighten the bounds.

4 Comparing Cross-Validation and Theoretical λ

4.1 Theoretical Choice of λ

The theorem in Tshibirani et al. provides a data-independent λ under specific assumptions:

1. **Assumption** : The design matrix $\mathbf{X}$ satisfies the restricted eigenvalue condition over $\mathcal{C}(S; 3)$
2. **Guarantee** : For $\lambda_N \geq \frac{2\|\mathbf{X}^T\mathbf{w}\|_\infty}{N}$, the Lasso estimate $\hat{\beta}$ satisfies:

   $$\|\hat{\beta} - \beta^*\|_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{N}}\lambda_N$$

   where $\gamma > 0$ is the RE parameter, k is the sparsity level, and $\mathbf{w}$ is the noise vector.

4.2 Cross Validation Choice of λ

Cross Validation (CV) selects λ empirically by:

1. Partitioning data into training/validation sets.
2. Choosing λ that minimizes prediction error on the validation set.

4.3 Advantages of CV over Theoretical λ

1. **Adaptability to Data** :
   (a) Theorem 11.1 assumes the RE condition and knowledge of γ, which are often unknown in practice. CV requires no such assumptions and adapts λ to the observed data.
   (b) The theoretical λ is conservative (designed for worst-case guarantees), while CV optimizes λ for the specific dataset.

2. **Handling Model Mismatch** : If assumptions like exact sparsity or RE condition are violated, the theoretical λ may over/under-regularize. CV automatically adjusts to the true signal structure.

3. **Practical Performance** : The theoretical λ prioritizes statistical guarantees (error bounds), often leading to over-regularization. CV balances prediction accuracy and model complexity, typically achieving lower prediction error in practice.

4. **Noise Adaptation** : Theorem 11.1 requires $\lambda_N \propto \|\mathbf{X}^T\mathbf{w}\|_\infty$, which depends on unobserved noise $\mathbf{w}$. CV bypasses this by directly measuring validation error.

4.4 Limitations of Cross Validation

1. **Computational Cost** : CV requires fitting the model multiple times, unlike the closed-form λ in Theorem 11.1.

2. **Theoretical Guarantees** : CV lacks finite-sample error bounds, whereas Theorem 11.1 provides explicit guarantees under its assumptions.

5 Using Mozorov's Criterion for λ Selection

According to the Mozorov criterion, the optimal value of lambda is chosen to minimize the following expression:

$$\lambda^* = \arg\min_{\lambda} |\|\mathbf{y} - \Phi\hat{\mathbf{x}}\|_2^2 - M * \sigma^2|$$

where,

1. M is the number of measurements

2. σ^2 is the noise variance

3. $\mathbf{y}$ is the observed signal

4. Φ is the sensing matrix

5.1 Advantages of Mozorov's Criterion

1. Provides a rigourous criterion based on the noise level which aligns well with the problems where the noise variance is known.

2. The method is less computationally expensive than cross-validation which needs the data to be split into training and validation sets and solving multiple optimization problems. Since this method uses directly observed data making it faster

3. The method is more stable where data is scarce or noisy unlike cross validation which can be sensitive to the choice of training and validation splits.

5.2 Disadvantages of Mozorov's Criterion

1. Requires a more accurate estimation of the noise variance σ^2 which can be difficult in practice. Otherwise the chosen λ may not be optimal leading to either overfitting or insufficient regularisation.

2. The method does not validate against unseen data so the chosen λ may not generalize well to new data.
3. In supervised learning problems, the goal is prediction accuracy rather than reconstruction, cross validation is preferred since it directly minimises the prediction error.
4. If the assumed noise model does not match the true noise characteristics, the discrepancy principle may fail to choose the optimal λ.

5.3 Experimental Results

Similar to Cross Validation, we implemented a MATLAB code to find the optimal value of λ using the Mozorov criterion. The results are shown in figure 3. The optimal value of λ is around $5 - 10$ which is similar to the values obtained from cross validation. This shows that the Mozorov criterion is a good alternative to cross validation for choosing λ in the given setup.

6 K-Fold Cross Validation

The symbol K in the paper refers to the **Number of Folds** in the K-fold cross validation. This is the number of partitions the dataset is divided into for training and validation. The parameter λ is chosen by minimizing the average prediction error over the K folds. The process is used as it balances the efficiency and reliability of the choice of λ.

6.1 Theorem 4.1 : Cross Validated LASSO

For any $\alpha \in (0,1)$,

$$\|\hat{\beta}(\hat{\lambda}) - \beta\|_{2,n} \leq \sqrt{\frac{Cs\log(p/\alpha)}{n}} \times \sqrt{\log(pn) + s^{-1}\log^{r+1} n}$$

with probability at least $1 - \alpha - Cn^{-c}$, where $c, C > 0$ are constants depending only on c_1, C_1, K, a, q, and r.
The paper also claims that the cross validated LASSO estimator has the fastest possible rate of convergence in the prediction norm upto a small factor.

6.2 Key Differences between the Bounds

1. Choice of Penalty Parameter (λ)
   (a) Tibshirani et al. provides a theoretically determined data independent λ that doesn't depend on the noise structure. This is not adaptive and assumes prior knowledge of the noise properties.
   (b) The Cross Validation Paper uses data driven K Fold Cross Validation to select λ. This does not require explicit theoretical constraints or prior knowledge of noise. The method is more practical but introduces additional variability.
2. Assumptions Made
   (a) Tibshirani et al. relies on the Restricted Eigenvalue condition on the design matrix X ensuring invertibility over sparse vectors. This is a strong geometric assumption.
   (b) Cross Validation Paper imposes weaker assumptions like sparsity, moment conditions on covariates like bounded sparse eigenvalues and smooth noise transformations. It avoids the RE condition making the results more generalisable.

3. Error Bounds

   (a) Tibshirani et al. achieves an L2 error bound scaling as $\mathcal{O}\left(\frac{k\lambda}{\gamma}\right)$ where γ is the RE parameter. This is rate optimal under the RE condition.

   (b) The paper derives a prediction error bound of $\mathcal{O}\left(\sqrt{\frac{s\log p}{n}} \cdot \sqrt{\log pn}\right)$ which is nearly optimal but includes additional logarithmic terms. These terms arise because of the adaptive nature of the method and the high dimensional setting.

4. Practicality vs Theoretical Tightness

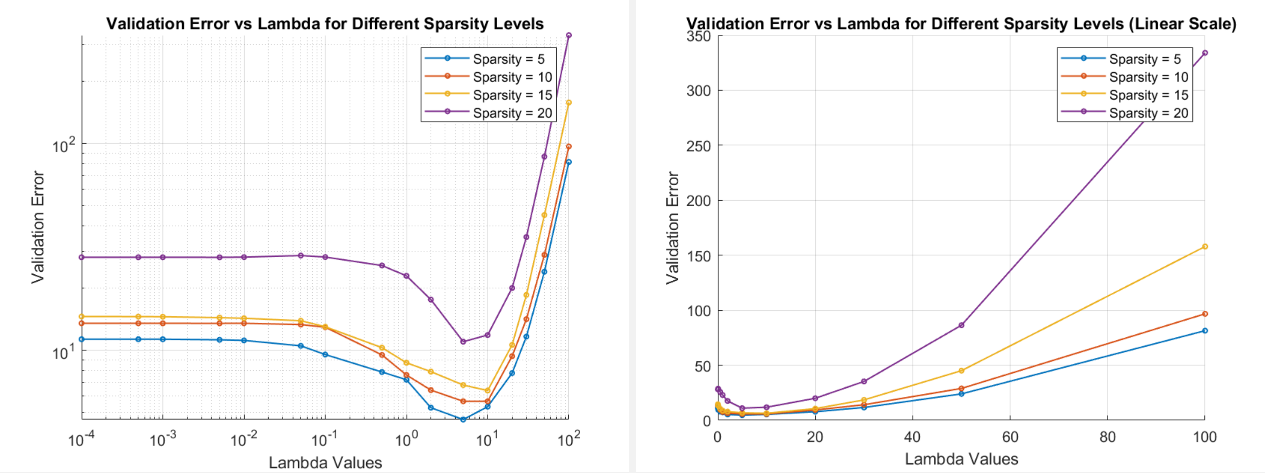
   (a) Tibshirani et al. provides a theoretically tight bound but requires knowledge of noise properties and RE conditions which are hard to verify in practice.

   (b) Cross Validation sacrifices tightness for adaptivity and broader applicability, accomodating unknown noise and avoiding RE conditions. This makes it more practical for real-world applications.

In conclusion, cross-validated LASSO achieves near-optimal error bounds with weaker assumptions and practical λ selection at the cost of logarithmic factors in the error bound. In contrast, Tibshirani's bound is theoretically tighter but relies on stronger assumptions and non-adaptive λ selection. The choice between the two methods depends on the specific problem context and the available information about the data.

Sparsity	Validation Error	RMSE	Mozorov Criterion
5	5	5	5
10	10	5	10
15	10	5	10
20	5	5	10

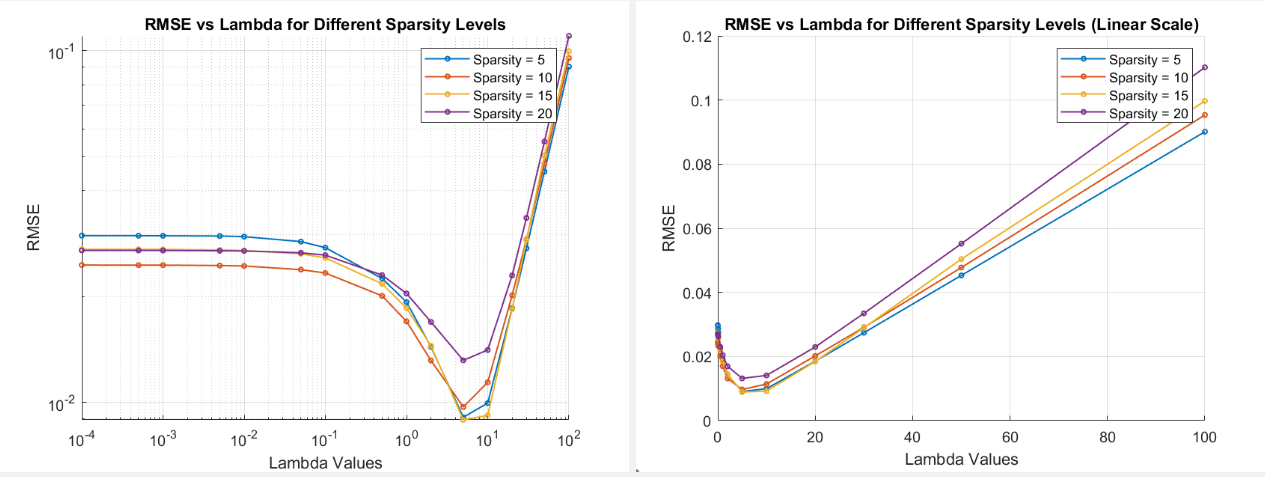
Table 1: Optimal λ for different values of sparsity



(a) Log Scale

(b) Linear Scale

Figure 1: Validation Error vs. λ for different values of sparsity



(a) Log Scale

(b) Linear Scale

Figure 2: RMSE vs. λ for different values of sparsity

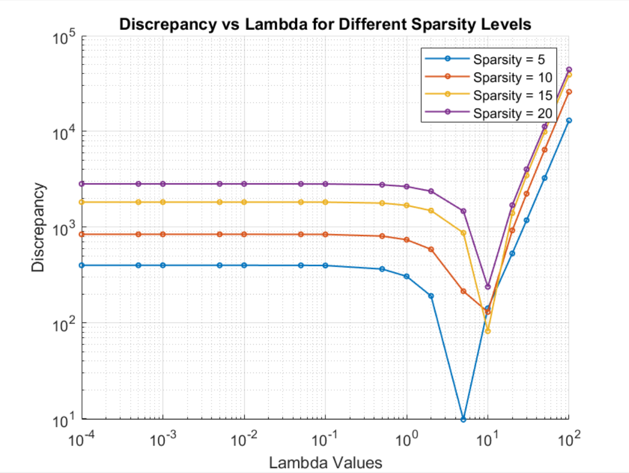


Figure 3: Optimal λ using the Mozorov Criterion