Report | HW 4

CS754 - Advanced Image Processing

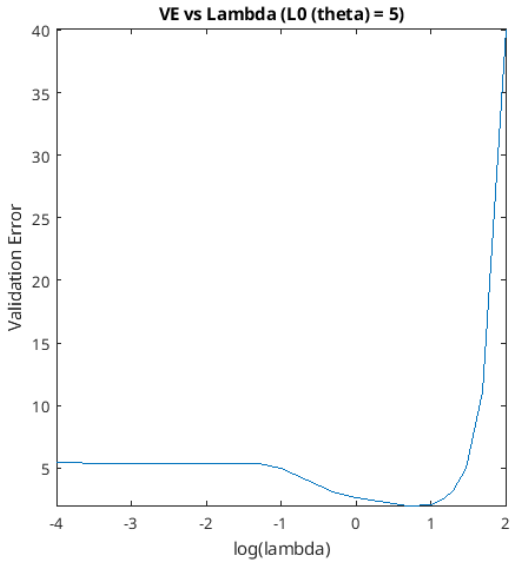
Omkar Shirpure (22B0910)
Krish Rakholiya (22B0927)
Suryansh Patidar (22B1036)

Contents

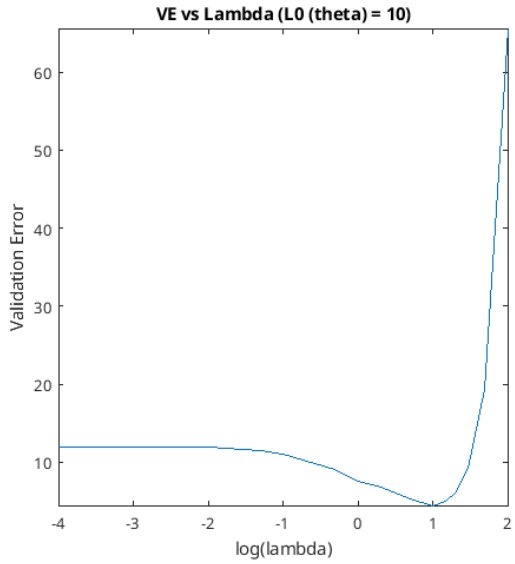
Declaration: The work submitted is our own, and we have adhered to the principles of academic honesty while completing and submitting this work. We have not referred to any unauthorized sources, and we have not used generative AI tools for the work submitted here.

1 Q1

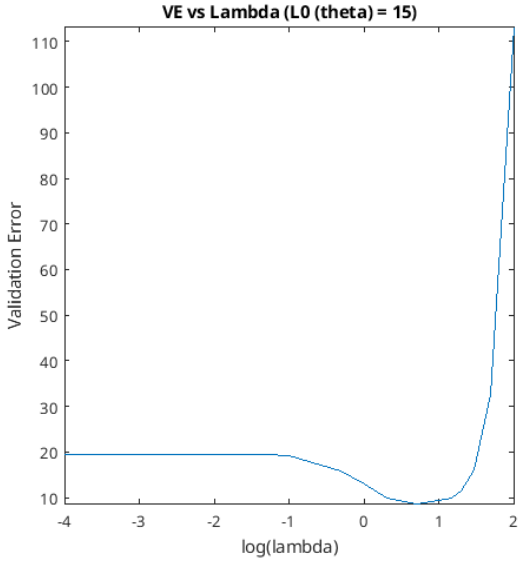
(a)



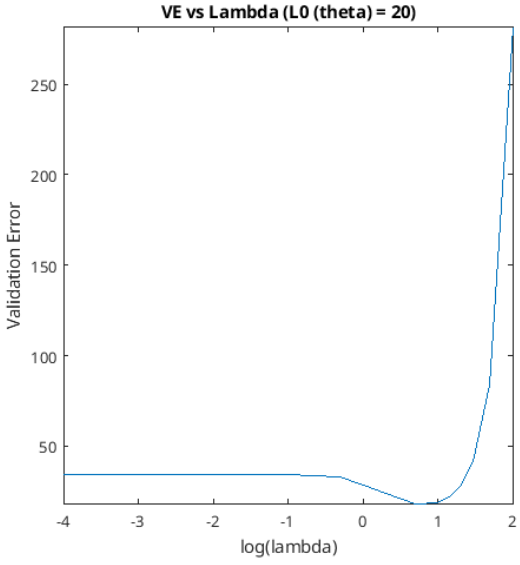
(a) $\|\boldsymbol{\theta}\|_0 = 5$, $\lambda_{\text{min}} = 5.0$



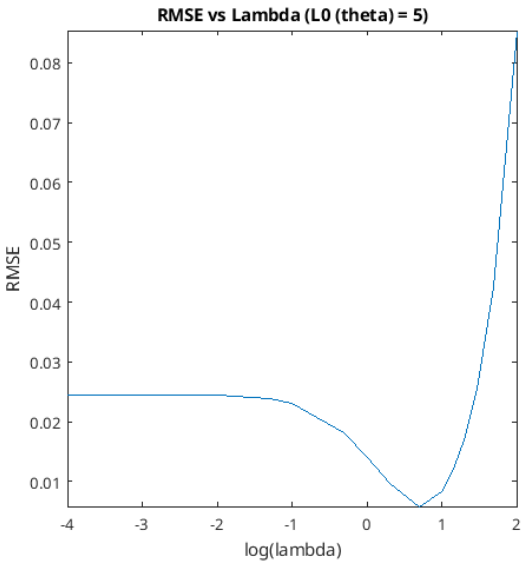
(b) $\|\boldsymbol{\theta}\|_0 = 10$, $\lambda_{\text{min}} = 10.0$



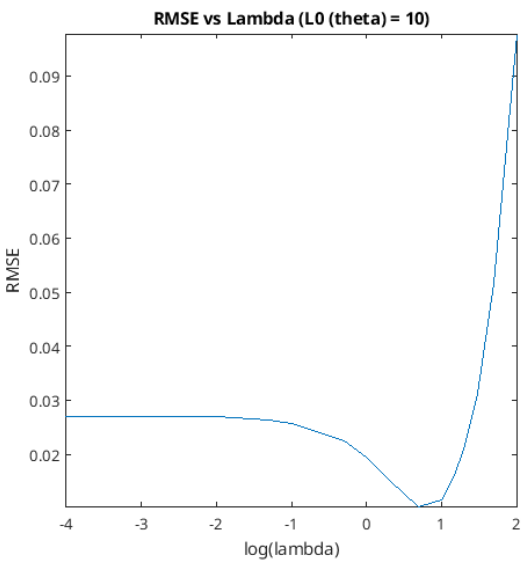
(a) $\|\boldsymbol{\theta}\|_0 = 15$, $\lambda_{\text{min}} = 5.0$



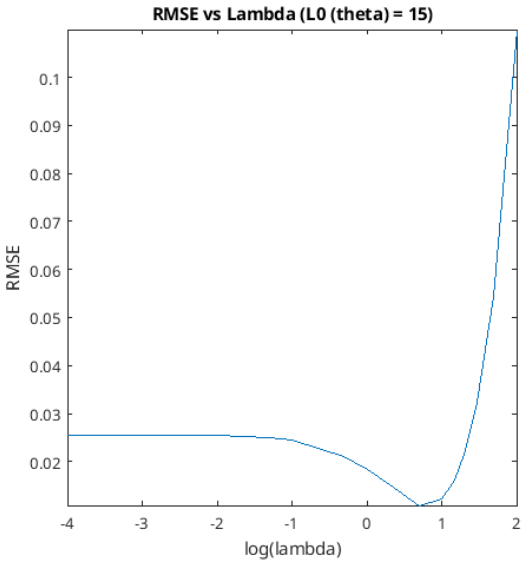
(b) $\|\boldsymbol{\theta}\|_0 = 20$, $\lambda_{\text{min}} = 5.0$



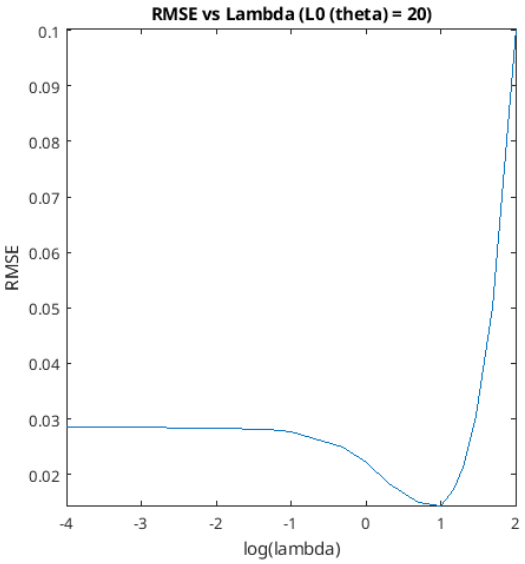
(a) $\|\boldsymbol{\theta}\|_0 = 5$, $\lambda_{\text{min}} = 5.0$



(b) $\|\boldsymbol{\theta}\|_0 = 10$, $\lambda_{\text{min}} = 5.0$



(a) $\|\boldsymbol{\theta}\|_0 = 15$, $\lambda_{\text{min}} = 5.0$



(b) $\|\boldsymbol{\theta}\|_0 = 20$, $\lambda_{\text{min}} = 10.0$

The optimal values for λ agree fairly closely among the RMSE and validation error plots.

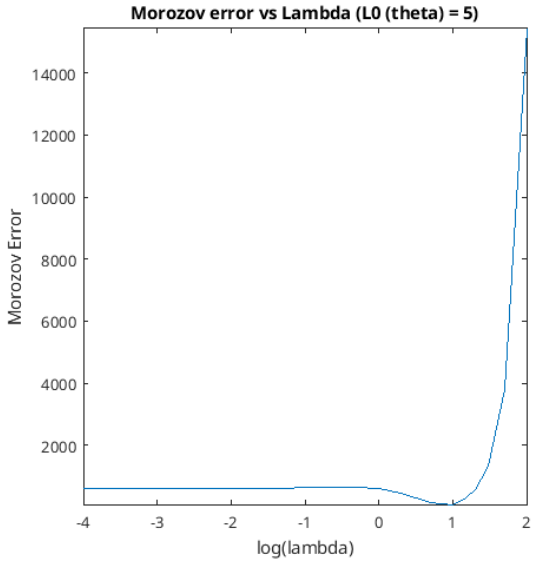
For the nature of the graphs, all graphs have a global point of minima for the error function at a certain optimal value of λ.

(b) If V and R were not disjoint but coincident sets, it could lead to biased results in the evaluation of the reconstruction algorithm. This is because the algorithm might perform well on the signals it has already seen during training, but it might not generalize well to unseen data. In such a scenario, the validation error would not accurately reflect the algorithm's true performance on unseen data because the algorithm may essentially be overfitting to the signals in the validation set.

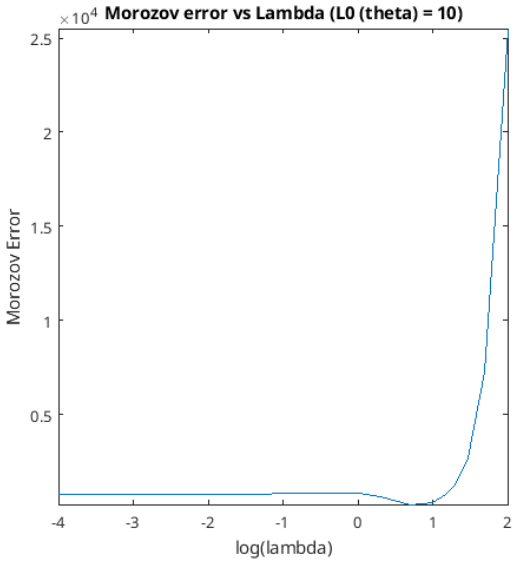
(c) Theorem 1 in the paper refers to the proxying ability of the validation error by essentially proving an upper and a lower bound on the MSE (ϵ_x) in terms of the cross-validation error ϵ_{cv} in equation (3).

(d) An advantage is that the theorem only gave a lower bound on the value of λ when the estimated and actual signals are close to each other, while the theorem for cross-validation in the previous part can be used to obtain both an upper bound and a lower bound on the values of λ.

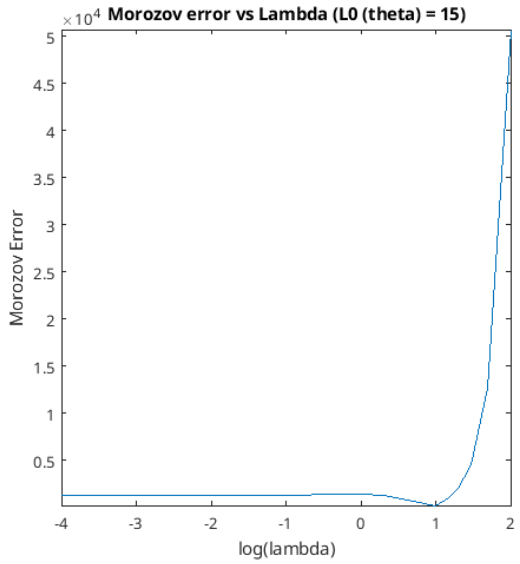
(e)



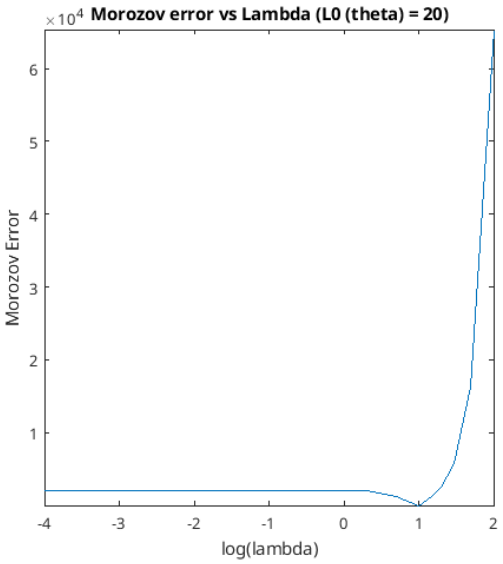
(a) $\|\boldsymbol{\theta}\|_0 = 5$, $\lambda_{\text{min}} = 10.0$



(b) $\|\boldsymbol{\theta}\|_0 = 10$, $\lambda_{\text{min}} = 5.0$



(a) $\|\boldsymbol{\theta}\|_0 = 15$, $\lambda_{\text{min}} = 10.0$



(b) $\|\boldsymbol{\theta}\|_0 = 20$, $\lambda_{\text{min}} = 10.0$

The advantage of this method is that it relies on the additional knowledge of the normally-unknown noise variance σ^2 to choose the optimal value of λ. This can be useful in practice when the noise variance is either known or can be estimated accurately. However, the disadvantage is that it may not always be possible to accurately estimate the noise variance, and the method may not perform well in such cases. On the other hand, cross-validation does not require any additional knowledge of the noise variance and can be used in a wider range of scenarios.

(f)

1. Meaning of the Symbol K in the Paper

In the paper *"On cross-validated Lasso in high dimensions"*, the symbol K denotes the **number of folds in K-fold cross-validation**.

- In K-fold cross-validation, the dataset is divided into K parts.
- Each fold is used once as the validation set, while the remaining $K-1$ folds are used for training.
- The paper establishes theoretical guarantees for Lasso when the regularization parameter λ is selected via cross-validation, uniformly over all $K \in \{2, \ldots, K_{\text{max}}\}$.

2. Comparison of Bounds

Theorem 4.1 from the Paper

Theorem 4.1 provides an upper bound on the prediction error of the Lasso estimator when the regularization parameter is selected via K-fold cross-validation. The bound has the general form:

$$\|X\hat{\beta}^{CV} - X\beta^*\|^2 \leq C \cdot \text{Oracle Error} + \text{Extra Terms}$$

Key points:

- $\hat{\beta}^{CV}$: Lasso estimate with λ selected by cross-validation.
- The result holds uniformly over all $K \in \{2, \ldots, K_{\text{max}}\}$.
- There is an additional penalty due to the data-driven selection of λ.

Bounds from the Previous Assignment

From the book *Statistical Learning with Sparsity,* Theorem 11.1 provides bounds for Lasso under ideal conditions.

Sparse Signal: Equation (11.15)

$$\|X(\hat{\beta} - \beta^*)\|_2^2 \leq C \cdot \frac{s \log p}{n}$$

Compressible Signal: Equation (11.16)

$$\|X(\hat{\beta} - \beta^*)\|_2^2 \leq C' \cdot \left(\frac{\log p}{n}\right)^{\frac{2r}{1+2r}}$$

Where:

- s: sparsity level of the true parameter β^*
- p: number of predictors
- n: number of samples
- r: decay rate of coefficients (for compressible signals)
- These bounds assume a theoretically optimal choice of λ, not selected via CV.
- Theorem 4.1 shows that Lasso with cross-validated λ performs nearly as well as an oracle version.

- Although the assignment bounds are theoretically tighter, they assume prior knowledge of the optimal λ, which is not available in practice.
- The paper provides theoretical support for the common practical approach of selecting λ via cross-validation.

2 Q2

(a) Probability of Getting a New Coupon on the jth Trial

We assume:

- There are n different coupons.
- Each trial is independent and samples uniformly at random with replacement.

On the jth trial, assuming the previous $j-1$ coupons were all distinct, the number of remaining new coupons is $n-(j-1)$. Thus, the probability that the jth coupon is new is:

$$q_j = \frac{n-j+1}{n}, \quad \text{with } q_1 = 1$$

(b) Probability that First Head Appears on the kth Trial

Let Y be the trial number on which the first head appears. Assuming head probability is q, the probability of $k-1$ tails followed by a head is:

$$\boxed{P(Y=k) = (1-q)^{k-1}q}$$

This is the geometric distribution.

(c) Expectation of Y

The expectation of the geometric random variable Y is:

$$\mathbb{E}[Y] = \sum_{k=1}^{\infty} k(1-q)^{k-1}q$$

Starting with the geometric series:

$$S = \sum_{k=0}^{\infty} p^k = \frac{1}{1-p}, \quad |p| < 1$$

Differentiating both sides with respect to p:

$$\frac{dS}{dp} = \sum_{k=0}^{\infty} kp^{k-1} = \frac{1}{(1-p)^2}$$

Shifting the index:

$$\sum_{k=1}^{\infty} kp^{k-1} = \frac{1}{(1-p)^2}$$

Substituting $p = 1-q$:

$$\mathbb{E}[Y] = q \cdot \frac{1}{(1-(1-q))^2} = q \cdot \frac{1}{q^2} = \boxed{\frac{1}{q}}$$

(d) Variance of Y

The variance is:

$$\mathsf{Var}(Y) = \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2$$

First compute $\mathbb{E}[Y^2]$:

$$\mathbb{E}[Y^2] = \sum_{k=1}^{\infty} k^2 (1-q)^{k-1} q$$

From the previous result:

$$\sum_{k=1}^{\infty} k p^{k-1} = \frac{1}{(1-p)^2}$$

Multiply by p:

$$\sum_{k=1}^{\infty} k p^{k} = \frac{p}{(1-p)^2}$$

Differentiate again:

$$\sum_{k=1}^{\infty} k^2 p^{k-1} = \frac{(1-p)^2 + 2p(1-p)}{(1-p)^4} = \frac{1+p}{(1-p)^3}$$

Substituting $p = 1-q$:

$$\mathbb{E}[Y^2] = q \cdot \frac{2-q}{q^3} = \frac{2-q}{q^2}$$

Finally:

$$\mathsf{Var}(Y) = \frac{2-q}{q^2} - \frac{1}{q^2} = \frac{1-q}{q^2}$$

$$\boxed{\mathsf{Var}(Y) = \frac{1-q}{q^2}}$$

(e) Expected Number of Trials to Collect All Coupons

Let Z_n be the number of trials needed to collect all n coupons. Let X_i be the number of trials needed to collect the ith new coupon after having seen $i-1$ distinct ones. Then:

$$X_i \sim \text{Geometric}\left(\frac{n-i+1}{n}\right) \Rightarrow \mathbb{E}[X_i] = \frac{n}{n-i+1}$$

Thus:

$$\mathbb{E}[Z_n] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{n}{n-i+1} = n \sum_{i=1}^{n} \frac{1}{i} = \boxed{nH_n}$$

where $H_n = \sum_{i=1}^{n} \frac{1}{i}$ is the nth harmonic number. Asymptotically:

$$H_n \approx \ln n + \gamma + \frac{1}{2n}$$

Variance Upper Bound

The variance of each X_i is:

$$\text{Var}(X_i) = \frac{1-q_i}{q_i^2} = \frac{n(n-i)}{(n-i+1)^2}$$

So:

$$\text{Var}(Z_n) = \sum_{i=1}^{n} \text{Var}(X_i) \leq n^2 \sum_{i=1}^{n} \frac{1}{i^2} < n^2 \cdot \frac{\pi^2}{6}$$

$$\boxed{\text{Var}(Z_n) < \frac{n^2\pi^2}{6}}$$

(f) Markov's Inequality

Markov's inequality gives:

$$P(Z_n \geq t) \leq \frac{\mathbb{E}[Z_n]}{t} = \frac{nH_n}{t}$$

$$\boxed{P(Z_n \geq t) \leq \frac{nH_n}{t}}$$

(g) Chebyshev's Inequality

Chebyshev's inequality states:

$$P(|Z_n - \mu| \geq \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

This gives a two-sided bound:

$$P(Z_n \leq \mu - \epsilon) + P(Z_n \geq \mu + \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

From this, we can extract the one-sided bound:

$$P(Z_n \geq \mu + \epsilon) \leq \frac{\sigma^2}{\epsilon^2}$$

Now, Let:

$$t = \mu + \epsilon \quad \Rightarrow \quad \epsilon = t - \mu$$

Then we get:

$$P(Z_n \geq t) \leq \frac{\mathsf{Var}(Z_n)}{(t-\mu)^2}$$

Using known values:

$$\mathsf{Var}(Z_n) \leq \frac{n^2\pi^2}{6}, \quad \mu = nH_n$$

Substitute into the bound:

$$P(Z_n \geq t) \leq \frac{n^2\pi^2/6}{(t - nH_n)^2}$$

Or more cleanly:

$$\boxed{P(Z_n \geq t) \leq \frac{n^2\pi^2}{6(t - nH_n)^2}}$$

3 Q3

Step 1: Measurement Model

The measurement matrix Y is formed by:

$$Y = \Phi M,$$

where Φ is an $m \times n_1$ random Gaussian matrix with entries drawn i.i.d. from $\mathcal{N}(0, 1)$.

Step 2: Rank Preservation

Random Gaussian matrices Φ preserve the rank of M with high probability:

$$\text{rank}(Y) = \text{rank}(M) = r,$$

provided:

- $m \geq r$,
- The entries of Φ are i.i.d. Gaussian (or satisfy the Restricted Isometry Property for low-rank matrices).

This property follows from the **Restricted Isometry Property (RIP)** for low-rank matrices.

Step 3: Determining r from Y

1. **Compute the SVD of Y**:
   $$Y = U\Sigma V^T,$$
   where Σ is a diagonal matrix of singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{\min(m,n_2)}$.
2. **Count Non-zero Singular Values**:
   - In noiseless settings, $\text{rank}(Y)$ equals the number of non-zero singular values.
   - Numerically, threshold tiny singular values (e.g., $\sigma_k > 10^{-10}$) to account for rounding errors.
3. **Result**: The count of significant singular values is the estimated rank r.

Key points to note:

- **Minimum** m: Ensure $m \geq r$. If $m < r$, $\text{rank}(Y) \leq m$, and r cannot be determined.
- **Noise Robustness**: For noisy measurements, use robust methods like Theorem 3 from or thresholding.
- **Efficiency**: For large Y, use randomized SVD to compute singular values.
- Random Gaussian matrices satisfy the **rank-RIP**, ensuring $\text{rank}(Y) = \text{rank}(M)$ with high probability if $m = O(r \max(n_1, n_2))$.
- The SVD of Y directly reveals r because Φ does not reduce the rank of M.

4 Q4

(a)

Since **x** has k non-zero elements and each A_{ij} is non-zero independently with probability γ, d_i counts the number of successes in k independent Bernoulli trials with success probability γ. Thus,

$$d_i \sim \text{Binomial}(k, \gamma)$$

(b)

To prove that $P(y_i = 0) = P(d_i = 0)$.

The i-th entry of **y** is:

$$y_i = \sum_{j=1}^{n} A_{ij} x_j$$

$y_i = 0$ occurs if and only if all $A_{ij}x_j = 0$. Since the non-zero A_{ij} are drawn from a continuous distribution, the probability that non-zero terms cancel exactly is zero. Therefore, $y_i = 0$ happens precisely when $A_{ij} = 0$ for all j where $x_j \neq 0$, which is the event $d_i = 0$. Hence,

$$P(y_i = 0) = P(d_i = 0)$$

(c)

From part (b), each y_i is non-zero with probability $1 - P(d_i = 0) = 1 - (1-\gamma)^k$. Since the y_i's are independent,

$$H \sim \text{Binomial}\big(m, 1 - (1-\gamma)^k\big)$$

(d)

From $P(d_i = 0) = (1-\gamma)^k$, taking logs gives:

$$k = \frac{\ln P(d_i = 0)}{\ln(1-\gamma)}$$

The MLE of $P(d_i = 0)$ is $\frac{m-H}{m}$, so:

$$\hat{k} = \frac{\ln\left(\frac{m-H}{m}\right)}{\ln(1-\gamma)}$$

(e)

$\hat{P} = \frac{m-H}{m}$ is a sample mean of m Bernoulli trials, so by the Central Limit Theorem, for large m, it is approximately Gaussian:

$$\hat{P} \sim \mathcal{N}\left((1-\gamma)^k, \frac{(1-\gamma)^k(1-(1-\gamma)^k)}{m}\right)$$

A q-confidence interval for $(1-\gamma)^k$ is:

$$\hat{P} \pm z_{(1+q)/2} \cdot \sqrt{\frac{\hat{P}(1-\hat{P})}{m}}$$

where $z_{(1+q)/2}$ is the $\frac{1+q}{2}$ quantile of $\mathcal{N}(0,1)$. Transforming to k:

$$\left[\frac{\ln\left(\hat{P} + z \cdot \text{SE}\right)}{\ln(1-\gamma)}, \frac{\ln\left(\hat{P} - z \cdot \text{SE}\right)}{\ln(1-\gamma)}\right]$$

(f)

With a prior $\pi(k)$, the Bayesian posterior is:

$$P(k|H) \propto P(H|k)\pi(k)$$

where $P(H|k)$ is the binomial likelihood from part (c). The MAP estimate is:

$$\hat{k}_{\text{MAP}} = \arg\max_k P(H|k)\pi(k)$$

Alternatively, we can compute the posterior mean or median, depending on $\pi(k)$.