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 4

Solution

We are given a vector $\mathbf{x} \in \mathbb{R}^n$ with k nonzero entries (unknown), a random sensing matrix $A \in \mathbb{R}^{m \times n}$, and the measurement vector $\mathbf{y} = A\mathbf{x} \in \mathbb{R}^m$. Each entry A_{ij} of A is

$$A_{ij} = \begin{cases} 0, & \text{with probability } 1-\gamma, \\ \mathcal{N}(0, \frac{1}{m\gamma}), & \text{with probability } \gamma, \end{cases}$$

independently for all i, j. We wish to estimate k (the number of nonzero entries of $\mathbf{x}$) directly from $\mathbf{y}$ and A.

(a) Distribution of d_i

Let d_i be the number of columns j for which both $A_{ij} \neq 0$ and $x_j \neq 0$. Assuming $\mathbf{x}$ has exactly k nonzero components, we have to find the distribution of d_i.
By definition, d_i only counts the columns j where $x_j \neq 0$. Since there are exactly k such nonzero entries of $\mathbf{x}$, only those k columns contribute.
For each nonzero x_j, the matrix entry A_{ij} is nonzero with probability γ. These events (over different j) are given to be independent. Hence d_i is the number of "successes" in k independent Bernoulli trials, each with success probability γ. Therefore

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

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

From part (a), $d_i = 0$ means that for *all* indices j with $x_j \neq 0$, the corresponding matrix entries A_{ij} are zero. In other words, none of the nonzero coordinates of $\mathbf{x}$ match with a nonzero entry in row i of A.
y_i is the ith component of $A\mathbf{x}$, i.e.

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

This sum will be zero *iff* every term $A_{ij}x_j$ is zero. Since $x_j = 0$ trivially contributes nothing, the only terms that matter are those with $x_j \neq 0$. If *all* of those corresponding A_{ij} are zero, the sum is necessarily zero which is exactly what yields $y_i = 0$. Thus,

$$\{y_i = 0\} \iff \{d_i = 0\},$$

and therefore

$$P(y_i = 0) \; = \; P(d_i = 0).$$

(c) Distribution of H = #(nonzero entries of y)

For each $i = 1, \dots, m$, define the indicator

$$I_i \; = \; \begin{cases} 1, & y_i \neq 0, \\ 0, & y_i = 0. \end{cases}$$

Then $H = \sum_{i=1}^{m} I_i$. From part (b), $y_i \neq 0$ is the same event as $d_i \neq 0$. Since $d_i \sim \text{Binomial}(k, \gamma)$,

$$P(d_i = 0) \; = \; (1-\gamma)^k, \quad \text{so} \quad P(d_i \neq 0) \; = \; 1 - (1-\gamma)^k.$$

Thus each I_i is a Bernoulli random variable with parameter $1 - (1-\gamma)^k$. Because A is drawn independently across rows and columns, the I_i are i.i.d. $\text{Bernoulli}\big(1 - (1-\gamma)^k\big)$. Therefore,

$$H \; = \; \sum_{i=1}^{m} I_i \; \sim \; \text{Binomial}\Big(m, \; 1 - (1-\gamma)^k\Big).$$

(d) MLE of k

From the derivation above, for each row i,

$$P\big(y_i = 0\big) \; = \; (1-\gamma)^k.$$

Equivalently,

$$(1-\gamma)^k \; = \; P\big(y_i = 0\big).$$

Instead of directly using $\mathbf{y}$, we can use the distribution of H. We know $H \sim \text{Binomial}\big(m, \; 1 - (1-\gamma)^k\big)$. The likelihood of observing H given k is

$$P(\mathbf{y}|k) = P(H|k) = Likelihood(k) \; = \; \binom{m}{H} \big[1 - (1-\gamma)^k\big]^H \big[(1-\gamma)^k\big]^{m-H}.$$

Taking log on both sides and ignoring the binomial term, we get

$$\ell(k) = H \ln\big[1 - (1-\gamma)^k\big] + (m-H)k\big[\ln(1-\gamma)\big]$$

We now compute the derivative of $\ell(k)$ with respect to k and set it to zero.

$$\frac{d\ell(k)}{dk} = 0 = H\Big(\frac{-(1-\gamma)^k \ln(1-\gamma)}{1-(1-\gamma)^k}\Big) + (m-H)\ln(1-\gamma)$$

We can factor out $\ln(1-\gamma)$:

$$\ln(1-\gamma) \left[-\frac{H\,(1-\gamma)^k}{1-(1-\gamma)^k} \; + \; (m-H) \right] \; = \; 0.$$

Since $\ln(1-\gamma) \neq 0$ for $0 < \gamma < 1$, the bracketed term must be zero:

$$-\frac{H\,(1-\gamma)^k}{1-(1-\gamma)^k} + (m-H) = 0.$$

$$(m-H) = \frac{H\,(1-\gamma)^k}{1-(1-\gamma)^k}.$$

$$(m-H)\left[1-(1-\gamma)^k\right] = H\,(1-\gamma)^k.$$

$$(m-H) - (m-H)\,(1-\gamma)^k = H\,(1-\gamma)^k.$$

$$(m-H) = m\,(1-\gamma)^k.$$

Therefore,

$$\frac{m-H}{m} = (1-\gamma)^k.$$

This also makes intuitive sense because $P(d_i = 0) = (1-\gamma)^k$, which empirically is $\frac{m-H}{m}$. Hence,

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

(e) Approximate confidence interval for k

Recall $\hat{P} = \frac{m-H}{m} = \frac{1}{m}\sum_{i=1}^{m} I\{y_i = 0\}$. Each indicator $I\{y_i = 0\}$ is a Bernoulli random variable with mean $(1-\gamma)^k$. We know that for large m, Binomial random variables fit the Normal distribution. By the Central Limit Theorem, $\hat{P}$ is approximately

$$\mathcal{N}\left((1-\gamma)^k,\ \frac{(1-\gamma)^k[1-(1-\gamma)^k]}{m}\right) \quad \text{for large } m.$$

Hence, it follows a normal distribution.
To create the confidence interval, we have $q = 1-\alpha$. A $(1-\alpha)$ confidence interval for $\hat{P}$ is

$$\hat{P} \pm z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{m}},$$

where $z_{\alpha/2}$ is the confidence level value. Note that the observed $\hat{P} = \frac{m-H}{m}$.
Now $\hat{k}$ is obtained from $\hat{P}$ by

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

Let $L(\hat{P}) = \hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{m}}$ and $U(\hat{P}) = \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(1-\hat{P})}{m}}$. Notice that $\ln(1-\gamma) < 0$, hence, we need to switch the endpoints. Hence, $L(\hat{k}) = \frac{\ln(U(\hat{P}))}{\ln(1-\gamma)}$ and $U(\hat{k}) = \frac{\ln(L(\hat{P}))}{\ln(1-\gamma)}$. Thus, we can say that

$$L(\hat{k}) \leq k \leq U(\hat{k}) \quad \text{with probability } q$$

(f) Incorporating a prior $\pi(k)$

Instead of simply using the likelihood to form an MLE, we use Bayes' rule to form the posterior distribution over k. Namely,

$$P(k \mid H) \;\propto\; \pi(k) \times P(H \mid k).$$

Here $P(H \mid k)$ is the binomial likelihood:

$$P(H \mid k) \;=\; \binom{m}{H} \left[1 - (1-\gamma)^k\right]^H \left[(1-\gamma)^k\right]^{m-H}.$$

We will now seek a MAP estimate, which maximizes $\pi(k)\, P(H \mid k)$ over k. This differs from the MLE, which uses only the likelihood, whereas MAP pulls the estimate towards the prior. Hence with a prior, the final estimate $\hat{k}$ is no longer simply $\ln\left(\frac{m-H}{m}\right) / \ln(1-\gamma)$. Instead, it is whichever value of k maximizes the posterior distribution given by

$$P(k \mid H) \propto \pi(k) \binom{m}{H} \left[1 - (1-\gamma)^k\right]^H \left[(1-\gamma)^k\right]^{m-H}.$$