Assignment 4: CS 754, Advanced Image Processing

Due: 4th April before 11:55 pm

Remember the honor code while submitting this (and every other) assignment. All members of the group should work on and understand all parts of the assignment. We will adopt a zero-tolerance policy against any violation.

Submission instructions: You should ideally type out all the answers in Word (with the equation editor) or using Latex. In either case, prepare a pdf file. Create a single zip or rar file containing the report, code and sample outputs and name it as follows: A4-IdNumberOfFirstStudent-IdNumberOfSecondStudent.zip. (If you are doing the assignment alone, the name of the zip file is A4-IdNumber.zip). Upload the file on moodle BEFORE 11:55 pm on the due date. The cutoff is 10 am on 5th April after which no assignments will be accepted. Note that only one student per group should upload their work on moodle. Please preserve a copy of all your work until the end of the semester. If you have difficulties, please do not hesitate to seek help from me.

Instructions for Coding Questions

1. Make a subfolder in the submission folder. Name the folder 'media'.
2. The directory structure should look like :

```
A4-<Roll_No_1>-<Roll_No_2>-<Roll_No_3>
    |
    |________media
    |________<other_file_1>
    |________<other_file_2>
    |________------------
    |________------------
    |________<other_file_n>
```

3. Read ANY image/video in ANY code from this folder(media) itself.
4. ALL the images/videos required for ANY code should be present in the folder 'media' itself, if your final compressed submission folder size DOES NOT EXCEED THE MOODLE SIZE LIMIT.
5. The TAs will copy all the images/video to the folder 'media' at the time of evaluation, if your final compressed submission folder DOES EXCEED THE MOODLE SIZE LIMIT. In this case leave the 'media' folder blank.
6. Please ensure that all the codes run at the click of a single go (RUN button) in MATLAB.
7. Please ensure that all the asked result images/videos, plots and graphs pop up at the click of a single go (RUN button) in MATLAB, while running the corresponding code for any question.
8. The result images/videos, plots and graphs should match those present in the report.

Questions

1. This question addresses a very practical implementation concern. Consider a signal $\boldsymbol{x}$ which is sparse in the 1D-DCT basis $\boldsymbol{\Psi} \in \mathbb{R}^{n \times n}$ and contains n elements. Let us suppose that the signal is compressively sensed in the form $\boldsymbol{y} = \boldsymbol{\Phi x} + \boldsymbol{\eta} = \boldsymbol{\Phi \Psi \theta} + \boldsymbol{\eta}$ where $\boldsymbol{y}$, the measurement vector, has m elements and $\boldsymbol{\Phi}$ is the $m \times n$ sensing matrix. Also $\boldsymbol{\theta}$ is a sparse vector of n coefficients. Here $\boldsymbol{\eta}$ is a vector of noise values that are distributed by $\mathcal{N}(0, \sigma^2)$. One way to recover $\boldsymbol{\theta}$ (and thereby also $\boldsymbol{x}$) from $\boldsymbol{y}, \boldsymbol{\Phi}$ is to solve the LASSO problem, based on minimizing $J(\boldsymbol{\theta}) \triangleq \|\boldsymbol{y} - \boldsymbol{\Phi \Psi \theta}\|^2 + \lambda \|\boldsymbol{\theta}\|_1$. A crucial issue is to how to choose λ. One purely data-driven technique is called cross-validation. In this technique, out of the m measurements, a random subset of (say) 90 percent of the measurements is called the reconstruction set $\mathcal{R}$, and the remaining measurements constitute the validation set $\mathcal{V}$. Thus $\mathcal{V}$ and $\mathcal{R}$ are always disjoint sets. The signal $\boldsymbol{x}$ is reconstructed using measurements only from $\mathcal{R}$ (and thus only the corresponding rows of $\boldsymbol{\Phi}$) using one out of many different values of λ chosen from a set Λ. Let the estimate using the g^{th} value from Λ be denoted $\hat{\boldsymbol{x}}_{\boldsymbol{g}}$. The corresponding validation error is computed using $VE(g) \triangleq \sum_{i \in \mathcal{V}} (y_i - \boldsymbol{\Phi^i} \hat{\boldsymbol{x}}_{\boldsymbol{g}})^2 / |\mathcal{V}|$. The value of λ for which the validation error is the least is chosen to be the optimal value of λ. Your job is to implement this technique for the case when $n = 500, m = 300, \|\boldsymbol{\theta}\|_0 \in \{5, 10, 15, 20\}, \sigma = 0.025 \times \sum_{i=1}^{m} |\boldsymbol{\Phi^i x}| / m$. Choose $\Lambda = \{0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, 15, 20, 30, 50, 100\}$. Draw the non-zero elements of $\boldsymbol{\theta}$ at randomly chosen location, and let their values be drawn randomly from Uniform$(0, 1000)$. The sensing matrix $\boldsymbol{\Phi}$ should be drawn from $\pm 1/\sqrt{m}$ Bernoulli with probability of $+1/\sqrt{m}$ being 0.5. Now do as follows. Use the popular CVX package (MATLAB version) for implementing the LASSO (or you may use your own previous ISTA code).

   (a) Plot a graph of VE versus the logarithm of the values in Λ for each value of $\|\boldsymbol{\theta}\|_0$. Also plot a graph of the RMSE versus the logarithm of the values in Λ, where RMSE is given by $\|\hat{\boldsymbol{x}}_{\boldsymbol{g}} - \boldsymbol{x}\|_2 / \|\boldsymbol{x}\|_2$. Comment on the plots. Do the optimal values of λ from the two plots agree? (Also see the last question in this list).

   (b) What would happen if $\mathcal{V}$ and $\mathcal{R}$ were not disjoint but coincident sets?

   (c) The validation error is actually a proxy for actual mean squared error. Note that you can never determine the mean squared error since the ground truth $\boldsymbol{x}$ is unknown in an actual application. Which theorem/lemma from the paper `https://ieeexplore.ieee.org/document/6854225` (On the theoretical analysis of cross-validation in compressed sensing) refers to this proxying ability? Explain how.

   (d) In your previous assignment, there was a theorem from the book by Tibshirani and others which gave you a certain value of λ. What is the advantage of this cross-validation method compared to the choice of λ using that theorem? Explain.

   (e) A curious student proposes the following method to choose λ: Pick the value from Λ for which $\|\boldsymbol{y} - \boldsymbol{\Phi} \hat{\boldsymbol{x}}_{\boldsymbol{g}}\|_2^2$ is the closest possible to $m\sigma^2$. This technique is motivated by the very definition of variance, and is often called Morozov's discrepancy principle. Implement this method as well, and plot a graph of RMSE and $|\|\boldsymbol{y} - \boldsymbol{\Phi} \hat{\boldsymbol{x}}_{\boldsymbol{g}}\|_2^2 - m\sigma^2|$ versus $\log \lambda$. What are the advantages and disadvantages of this method as compared to cross-validation?

   (f) Read the paper 'On cross-validated Lasso in high dimensions' published in the Annals of Statistics. The paper can be found at `https://projecteuclid.org/journals/annals-of-statistics/volume-49/issue-3/On-cross-validated-Lasso-in-high-dimensions/10.1214/20-AOS2000.full`. What is the meaning of the symbol K in the paper? How do the bounds in theorem 4.1 of this paper compare to the bounds of the LASSO that you studied in the previous assignment? [8+3+3+3+(5+5)+8=35 points]

   Part (a): A sample of plots of VE and RMSE versus $\log \lambda$ are in the homework folder ('crossval.png'). The plots show that the optimal value of λ in terms of least validation error is close to the one in terms of least RMSE, though they may not be exactly equal. In both cases, the RMSE/VE curves decrease with increase in λ, then dip down to a minimum and then increase w.r.t. λ. **Marking scheme:** 5 points for correct shape of the plots as output by the code. 3 points for the aforementioned comments on the plots.

   Part (b): It is important to have R and V that are disjoint from one another. Suppose you instead had $R = V$. Since the problem is inherently ill-posed without a sparsity constraint, you could find some $\boldsymbol{x}$ such

that $\sum_{i \in R}(y_i - \boldsymbol{\Phi^i x})^2$ is zero or close to zero. This is easy to do by choosing some vector from the nullspace of $\boldsymbol{\Phi_R}$. If V and R are the same, this will produce a low validation error, but such an estimate of $\boldsymbol{x}$ will be a very bad one as no sparsity prior has been used, and there are no theoretical guarantees for such an estimate. On the other hand, if V is disjoint from R, the validation error will likely be very high, as is also inferred from theorem 1 (see solution to part (c) as well). **Marking scheme:** 4 points for correct explanation – it must show why a fully overlapping V and R will be problematic.

Part (c): Theorem 1 from this paper refers to the proxying ability. It clearly gives a confidence interval for ε_x (the true MSE) in terms of ϵ_{cv} (the cross-validation error), σ_n (the noise standard deviation) and m_{cv} (the number of measurements used for cross-validation). This shows that ϵ_{cv} gives us some indication of the value of ε_x. **Marking scheme:** 4 points for identifying the theorem and explaining why it is useful. The explanation is a must for getting any points.

Part (d): That theorem indeed gave a lower bound for the value of λ. This derived lower bound is sufficient for the upper bounds on the estimation error, i.e. $\|\boldsymbol{x} - \hat{\boldsymbol{x}_\lambda}\|_2$ to hold. Larger values of λ produce looser upper bounds, as per that theorem. However, for specific instances, the value of λ predicted from that theorem need not produce the best MSE, nor do we have any confidence interval for the MSE. The present cross-validation approach instead gives us such a confidence interval. **Marking scheme:** 4 points for correct explanation.

Part (e): **Marking scheme:** 5 points for implementation of the discrepancy principle. 5 points for discussion of relative merits and demerits w.r.t. CV as follows (2.5 for merits, 2.5 for demerits): The DP does not require a split of the measurements into disjoint sets R and V, and thus does not lose out on measurements. However unlike CV, the DP requires knowledge of the noise variance, and the choice of the term $m\sigma^2$ is somewhat ad hoc. For example, why not $1.2m\sigma^2$?

Part (f): See theorem 4.1 from the Annals of Statistics paper (AoS). The paper derives error bounds on $\|\boldsymbol{\beta^*} - \hat{\beta}(\hat{\lambda})\|_2$, where $\hat{\beta}(\hat{\lambda})$ is the parameter chosen by K-fold cross-validation – see Theorem 4.1. As opposed to this, the LASSO bounds in the book by Tibshirani (equation 11.15) merely assume that λ satisfies a certain statistical lower bound, and derive an error bound for that. The bound in the AoS paper may appear looser than the one in the book due to the extra term $\sqrt{\log(pn) + \log^{r+1}(n)/s}$. However the bounds in the book are true for zero-mean Gaussian noise and the choice of λ requires knowledge of the noise standard deviation σ. On the other hand, the cross-validation based bounds in the AoS paper have no such requirements.

2. Consider that there are n coupons, each of a different colour. Suppose we sample coupons uniformly at random and with replacement. Then answer the following questions from first principles. Do not merely quote results pertaining to any distribution other than Bernoulli. [3+3+3+3+(4+4 = 8)+2.5+2.5 = 25 points]

   (a) What is the probability q_j of getting a new coupon on the jth trial, assuming that new, unique coupons were obtained in all the previous trials? What is q_1?

   (b) If you were to toss a coin independently many times, what is the probability that the first head appears on the kth trial? Assume that the probability of getting a heads on any trial is q.

   (c) Let Y be the random variable which denotes the trial number on which the first head appears. Derive a formula for $E(Y)$ in terms of q.

   (d) Derive a formula for the variance of Y.

   (e) Let Z_n be a random variable denoting the number of trials by which each of the n different coupons were selected at least once. Applying result in previous parts, what is the expected value of Z_n? Derive an upper bound on the variance of Z_n. (You will need to use the following results: $\sum_{i=1}^{n} 1/i \approx \log n + \gamma + O(1/n)$ where $\gamma \approx 0.5772$ is a constant, and $\sum_{i=1}^{n} 1/i^2 < \sum_{i=1}^{\infty} 1/i^2 < \pi^2/6$).

   (f) Given the previous results, use Markov's inequality to upper bound $P(Z_n \geq t)$ for some value t.

   (g) Given the previous results, use Chebyshev's inequality to upper bound $P(Z_n \geq t)$ for some value t.

Solutions:

(a) $q_j = \dfrac{n - j + 1}{n}$ and $q_1 = 1$.

(b) The probability that the first head appears on the kth trial is $(1-q)^{k-1}q$.

(c) By definition of expectation, $E(Y) = \sum_{k=0}^{\infty} k(1-q)^{k-1} \times q = q\sum_{k=0}^{\infty} k(1-q)^{k-1}$.
This produces $E(Y) = -q\sum_{k=0}^{\infty} \frac{d}{dq}(1-q)^k = -q\frac{d}{dq}\sum_{k=0}^{\infty}(1-q)^k = -q\frac{d}{dq}\frac{1}{1-(1-q)} = \frac{1}{q}$.

(d) We have $Var(Y) = E(Y^2) - [E(Y)]^2$. Then $E(Y^2) = \sum_{k=0}^{\infty} k^2(1-q)^{k-1} \times q$. This gives us

$$E(Y^2) = q\sum_{k=0}^{\infty}(k^2-k+k)(1-q)^{k-1} = q\left(\sum_{k=0}^{\infty} k(k-1)(1-q)^{k-1} + \sum_{k=0}^{\infty} k(1-q)^{k-1}\right) \quad (1)$$

$$= q\left(\frac{d^2}{dq^2}\sum_{k=0}^{\infty}(1-q)^k + \frac{d}{dq}\sum_{k=0}^{\infty}(1-q)^k\right) \quad (2)$$

$$= q\left(\frac{d^2}{dq^2}(1/q) + \frac{d}{dq}(1/q)\right) \quad (3)$$

$$= q\left(\frac{2}{q^3} - \frac{1}{q^2}\right) = \frac{2}{q^2} - \frac{1}{q}. \quad (4)$$

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

(e) Z_n is a random variable denoting the number of trials by which each of the n different coupons were selected at least once. We see that $Z_n = T_1 + T_2 + ... + T_n$ wheer T_i is the number of rounds to collect a coupon of the ith type after coupons of the previous $i-1$ types have been collected. From parts 1 and 2, we see that T_i is a geometric random variable with success probability $q_i = \frac{n-i+1}{n}$. Now, $E(Z_n) = \sum_{i=1}^{n} E(T_i) = \sum_{i=1}^{n} \frac{n}{n-i+1} = \frac{n}{n} + \frac{n}{n-1} + ... + \frac{n}{1} = n(1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n}) = nH_n$ where $H_n := (1 + \frac{1}{2} + \frac{1}{3} + ... + \frac{1}{n}) = \log n + \gamma + O(1/n)$. Hence $E(Z_n) = n\log n + \gamma n + O(1)$.
The variance of Z_n is given by $Var(Z_n) = \sum_{i=1}^{n} Var(T_i) = \sum_{i=1}^{n} \frac{1-q_i}{q_i^2} = \left(\frac{n^2}{n^2} + \frac{n^2}{(n-1)^2} + ... + \frac{n^2}{1^2}\right) - \left(\frac{n}{n} + \frac{n}{n-1} + ... + \frac{n}{1}\right) < \left(\frac{n^2}{n^2} + \frac{n^2}{(n-1)^2} + ... + \frac{n^2}{1^2}\right) = n^2\left(\frac{1}{n^2} + \frac{1}{(n-1)^2} + ... + \frac{1}{1^2}\right) = n^2\pi^2/6$. This is the upper bound for Z_n.

(f) Using Markov's inequality, we have $P(Z_n \geq t) \leq \frac{E(Z_n)}{t} = \frac{n\log n + \gamma n + O(1)}{t}$.

(g) Using Chebyshev's inequality, we have $P(|Z_n - E(Z_n)| \geq t') \leq \sigma^2/t'^2 < \frac{\pi^2 n^2}{6t'^2}$ where σ^2 is the variance of Z_n. Hence $P(Z_n \geq E(Z_n) + t') < \frac{\pi^2 n^2}{6t'^2}$. Hence $P(Z_n \geq t) < \frac{\pi^2 n^2}{6(t - E(Z_n))^2}$.

3. Consider a matrix $\boldsymbol{M}$ of size $n_1 \times n_2$ having low but unknown rank $r < \max(n_1, n_2)$. Suppose you observe noiseless compressive measurements in the form of a matrix $\boldsymbol{Y}$ of size $m \times n_2$ where every column of $\boldsymbol{Y}$ is obtained by taking dot-products of the corresponding column of $\boldsymbol{M}$ with m different vectors whose elements are drawn from a zero-mean Gaussian distribution. Note that $m < n_1, m < n_2$. How will you determine the rank of $\boldsymbol{M}$ given the measurements in $\boldsymbol{Y}$. [15 points]
Solution: Here we clearly have $\boldsymbol{Y} = \boldsymbol{\Phi M}$ where $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n_1}$ and $\boldsymbol{Y} \in \mathbb{R}^{m \times n_1}$. Using properties of the rank, we have the result from linear algebra that $\text{rank}(\boldsymbol{\Phi M}) \leq \min(\text{rank}(\boldsymbol{\Phi}, \text{rank}(\boldsymbol{M}))$. If $\boldsymbol{M}$ has a rank of $r < m$, then we have $\text{rank}(\boldsymbol{\Phi M}) \leq r$. Marking scheme: This answer is sufficient for 15 marks. However I have empirically observed that $\text{rank}(\boldsymbol{\Phi M}) = \text{rank}(\boldsymbol{M})$. This could be because we have the linear algebra result that $\text{rank}(\boldsymbol{\Phi M}) = \text{rank}(\boldsymbol{M})$ if $\boldsymbol{\Phi}$ is invertible (see `https://forms.gle/gkrba2tkNzeCB1Xm9`). Now given the dimensions of $\boldsymbol{\Phi}$, it is not even square and is rank-deficient. Nonetheless, $\boldsymbol{\Phi}$ it is still invertible for low-rank matrices $\boldsymbol{M}$ as per compressive low rank matrix recovery theory.

4. Let $\boldsymbol{x}$ be a real-valued vector of n elements. We have the relationship $\boldsymbol{y} = \boldsymbol{Ax}$ where $\boldsymbol{y}$ is a measurement vector with m elements, and $\boldsymbol{A}$ is a $m \times n$ sensing matrix, where $A_{ij} = 0$ with probability $1-\gamma$ and

$A_{ij} = \mathcal{N}(0, 1/m\gamma)$ with probability γ. All entries of $\boldsymbol{A}$ are drawn independently. Here $\gamma \in (0, 1)$. Note that we have no knowledge of k beforehand. This question seeks to explore a technique to estimate k directly from $\boldsymbol{y}$ and $\boldsymbol{A}$. To this end, answer the following questions: [4+4+4+4+5+4=25 points]

(a) Let d_i be the number of entries for which A_{ij} and x_j are both unequal to 0, where $1 \leq j \leq n$. What is the distribution of d_i, if the number of non-zero elements of $\boldsymbol{x}$ is k, since the entries of $\boldsymbol{A}$ are drawn independently?

(b) Prove that $P(y_i = 0) = P(d_i = 0)$.

(c) Let H be a random variable for the number of non-zero elements in $\boldsymbol{y}$. Then what is the distribution of H, if the number of non-zero elements of $\boldsymbol{x}$ is k?

(d) Express k in terms of $P(d_i = 0)$ and hence write the maximum likelihood estimate of k given $\boldsymbol{y}$.

(e) Let $\hat{P}$ be the estimate of $P(d_i = 0)$. Then $\hat{P}$ is an approximately Gaussian random variable. Explain why. Using this, state how you will provide a confidence interval for the true k using its estimate $\hat{k}$ derived so far. That is, you need to provide an interval of the form $L(\hat{k}) \leq k \leq U(\hat{k})$ with some probability q.

(f) Now consider that you had knowledge of some prior distribution $\pi(k)$ on k. How does your estimate of k now change?

Solution:

(a) Discarding pathological corner cases where the elements of $\boldsymbol{A^i}$ and $\boldsymbol{x}$ participate in a premeditated conspiracy where $\boldsymbol{A^i x} = 0$ even when both $\boldsymbol{A^i}$ and $\boldsymbol{x}$ contain non-zero elements, we have $d_i \sim \mathcal{B}(k, \gamma)$, i.e. a binomial distribution given k coin tosses each having success probability γ.

(b) $P(y_i = 0) = \sum_{j=0}^{n} P(y_i = 0 | d_i = j) P(d_i = j) = P(d_i = 0)$. This is because y_i cannot equal 0 if $d_i > 0$ as per the definition of d_i. Also, as per the definition of d_i, we must have $y_i = 0$ when $d_i = 0$.

(c) H is a binomial random variable obtained by adding m Bernoulli random variables with success probability $1 - q$ where $q = P(y_i = 0)$. Thus $H \sim \mathcal{B}(1 - q, m)$.

(d) From part (a), we see that $P(d_i = 0) = C(k, 0)\gamma^0(1 - \gamma)^k = (1 - \gamma)^k$. This is equal to $P(y_i = 0)$ from part (b). But $P(y_i = 0)$ can be estimated to be $1 - \|\boldsymbol{y}\|_0/m$. The ML estimate of k is thus given by $\hat{k} = \dfrac{\log P(y_i = 0)}{\log(1 - \gamma)} = \dfrac{\log(1 - \|\boldsymbol{y}\|_0/m)}{\log(1 - \gamma)}$. In practice, this may be a fraction, so there needs to be a rounding operation, i.e. $\hat{k} = \text{round}\left(\dfrac{\log(1 - \|\boldsymbol{y}\|_0/m)}{\log(1 - \gamma)}\right)$.

(e) $\hat{P} := 1 - \|\boldsymbol{y}\|_0/m$ is approximately Gaussian if m is large enough, because for large m, the Binomial distribution $\mathcal{B}(m, p)$ is approximated by a Gaussian of variance $mp(1 - p)$ and mean mp where p is the success probability of the Bernoulli trials. Hence the standard deviation of $\hat{P}$ is $\sqrt{p(1 - p)/m}$. We know that $\hat{k} = \dfrac{\log P(y_i = 0)}{\log(1 - \gamma)} = \dfrac{\log(1 - \hat{P})}{\log(1 - \gamma)}$. Let us use $f_0 = \|\boldsymbol{y}\|_0/m$. Using this Gaussian approximation for $\hat{P}$, we see that with probability 99%, we have

$$L_k \leq \hat{k} \leq U_k \tag{5}$$

where $L_k = \dfrac{\log\left(1 - f_0 - 3\sqrt{\dfrac{f_0(1 - f_0)}{m}}\right)}{\log(1 - \gamma)}$ and $U_k = \dfrac{\log\left(1 - f_0 + 3\sqrt{\dfrac{f_0(1 - f_0)}{m}}\right)}{\log(1 - \gamma)}$.

(f) Let the prior on k be denoted by $\pi(k)$. Then we find the MAP estimate instead of the MLE estimate for k. That is we find:

$$\hat{k} = \text{argmax}_k P(k|f_0) = \text{argmax} P(f_0|k)\pi(k) = \text{argmax}\mathcal{B}(1 - f_0, m)\pi(k). \tag{6}$$