College of Engineering
School of Aerospace Engineering

ISYE7750: Mathematical Foundations of Machine Learning

Homework 4

Professor:
Ashwin Pananjady
Gtech ISYE Professor

Student:
Tomoki Koike
AE MS Student

October 14, 2022

Table of Contents

I Problem One

Recall the bump basis $\{\phi_n(t)\}_{n=1}^N$ from Homework 2, Problem 3 (Linear approximation with "bump" functions), and its span $\mathcal{T}_N$ equipped with the standard inner product. The dual basis $\{\tilde{\phi}_n(t)\}_{n=1}^N$ can be used to find the sampling functions (reproducing kernel) for $\mathcal{T}_N$, as

$$f(\tau) = \sum_{n=1}^{N} \langle \mathbf{f}, \tilde{\phi}_n \rangle \phi_n(\tau) = \left\langle \mathbf{f}, \sum_{n=1}^{N} \phi_n(\tau)\tilde{\phi}_n \right\rangle = \langle \mathbf{f}, \mathbf{k}_\tau \rangle, \qquad \text{where } \mathbf{k}_\tau = \sum_{n=1}^{N} \phi_n(\tau)\tilde{\phi}_n.$$

(a) Fix $N = 10$ and compute the dual basis vectors of the bump basis from Homework 2, Problem 3. That is, find $\tilde{\phi}_1, \ldots, \tilde{\phi}_{10}$ so that if

$$f(t) = \sum_{n=1}^{10} \alpha_n \phi_n(t),$$

we can compute the $\{\alpha_n\}_{n=1}^N$ using

$$\alpha_n = \int_0^1 f(t)\tilde{\phi}_n(t)\, dt.$$

Turn in a plot of each of the ten $\tilde{\phi}_n(t)$.

(b) Take $N = 10$ and plot $k_\tau(t)$ as a function of t for $\tau = .371238$. Create an $\mathbf{f} \in \mathcal{T}_N$ by drawing the expansion coefficients α at random (`alpha = randn(N,1);` in MATLAB), and verify that $\langle \mathbf{f}, \mathbf{k}_\tau \rangle = f(\tau)$.

(c) Create an image of the kernel $k(s,t)$ for $(s,t) \in [0,1] \times [0,1]$ for the basis above — use at least a few hundred points for each of the arguments s and t. (In MATLAB you can display using `imagesc`.)

Solution

Question (a)

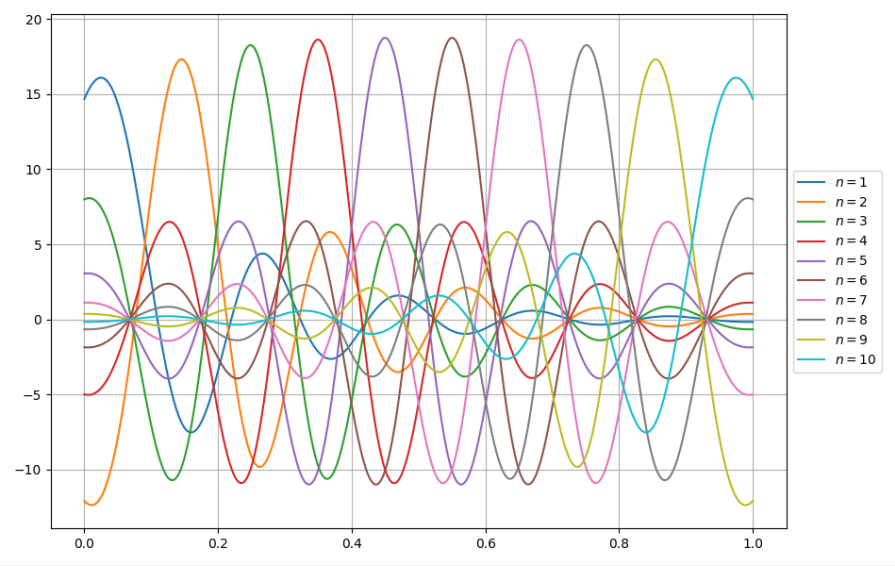

Figure 1: Dual basis $\tilde{\phi}(t)$ for $N = 10$.

The dual basis found from

$$\tilde{\phi}(t) = \sum_{l=1}^{N} H_{n,l}\phi_l(t),$$

where $\mathbf{H}$ is the inverse of the Gram matrix.

Question (b)

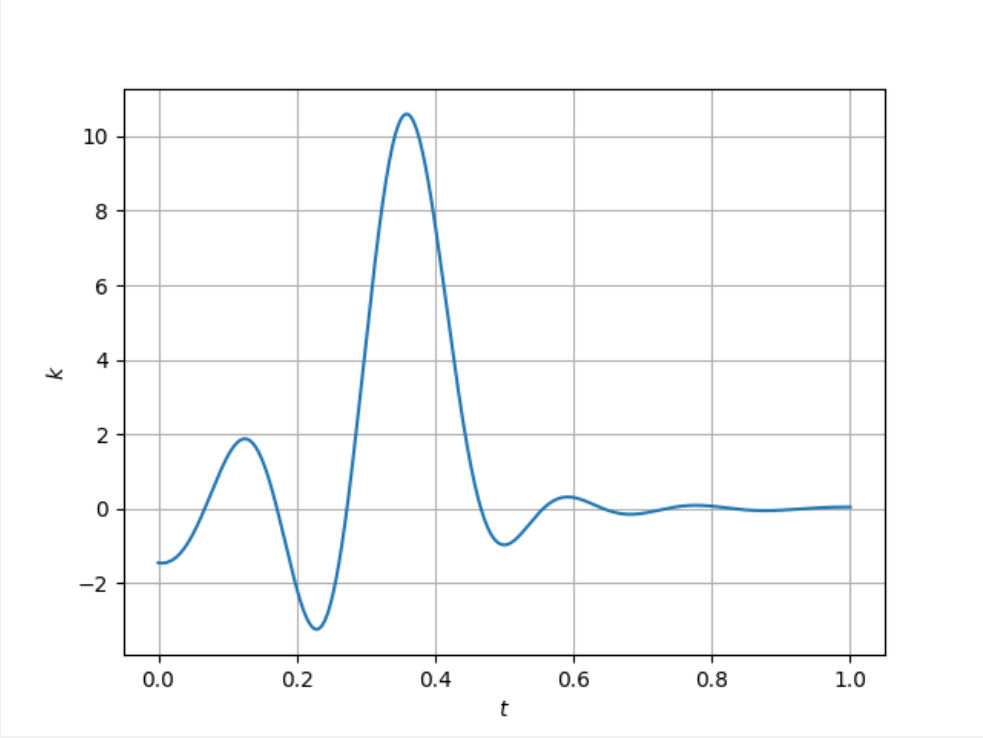

Figure 2: $k_\tau(t)$ vs t for $\tau = 0.371238$.

The randomly generated expansion coefficients are

$$\alpha = \begin{bmatrix} 0.1630 & 0.3583 & -0.3096 & -2.0965 & -0.2018 & -0.3279 & -0.7592 & -1.8327 & -0.4085 & -0.3872 \end{bmatrix}$$

With these coefficients we have

$$f(\tau) = -2.1948 \quad \text{and} \quad \langle \mathbf{f}, \mathbf{k}_\tau \rangle = -2.1947 \tag{I.1}$$

Thus we can verify that the kernel holds the reproducing property.

Question (c)

For this problem we use 1000 points between $[0, 1]$ for t and τ. The kernel image is presented in the next page.

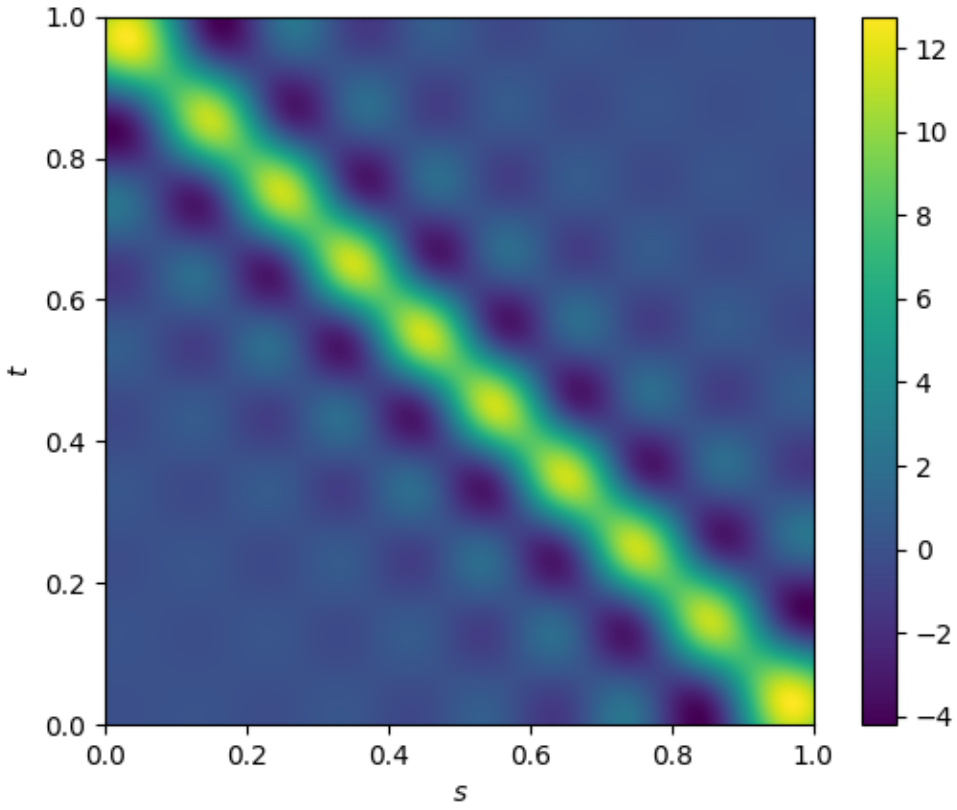

Figure 3: The kernel image $k(s,t)$.

The code producing all the plots are submitted separately.

II Problem Two

In this problem, we will solve a stylized regression problem using the data set `hw04p2_data.mat`. This file contains (noisy) samples of a function $f(t)$ for $t \in [0, 1]$. In fact, the data points were generated by sampling the function

$$f_{\text{true}}(t) = \frac{\sin(12(t + 0.2))}{t + 0.2}$$

at random locations then adding a random perturbation to the sample values. The sample locations are in the vector `T`, the sample values are in `y`. If you plot these, you will see that the samples are scattered more or less evenly across the interval. We are going to use kernel regression to form the estimate; in particular, we will use

$$k(s, t) = e^{-|t-s|^2/2\sigma^2}.$$

(a) Compute the kernel regression estimate with $\sigma = 1/10$ and $\delta = 0.004$. Plot your estimate $\hat{f}(t)$ overlaid on the data and $f_{\text{true}}(t)$. Compute the *sample error* (Also called the "training error")

$$\text{sample error} = \left(\sum_{m=1}^{M} |y_m - \hat{f}(t_m)|^2 \right)^{1/2},$$

and the *generalization error*

$$\text{generalization error} = \left(\int_0^1 |\hat{f}(t) - f_{\text{true}}(t)|^2 \right)^{1/2}$$

for your estimate. Comment on why this choice of σ was a good one.

(b) Repeat part (a) with $\sigma = 1/2, 1/5, 1/20, 1/50, 1/100, 1/200$, producing plots, sample errors, and generalization errors for your estimates for each σ. Comment on how the number of data points we see would affect the right choice of σ.

Solution

Question (a)

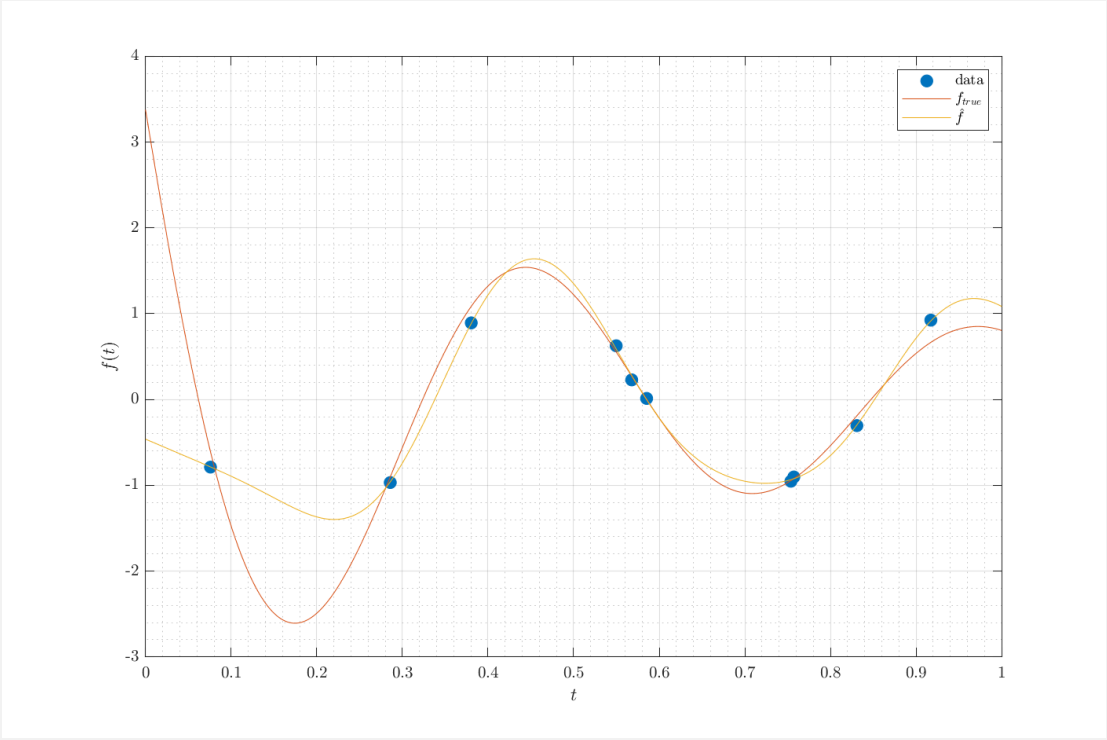

Figure 4: Plot of $\hat{f}$ and f_{true}.

The expansion coefficients computed for this kernel regression are as follows.

$$\hat{\alpha} = \begin{bmatrix} -1.9473 & 3.6189 & 9.4253 & 2.1429 & -2.1024 & 4.9473 & -4.7861 & 1.7431 & -12.7353 & -0.5708 \end{bmatrix}.$$

These were computed by following the procedure of computing

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \delta \mathbf{I})^{-1} \mathbf{y} \tag{II.1}$$

and

$$\hat{\mathbf{f}}(\boldsymbol{\tau}) = \sum_{m=1}^{M} \hat{\alpha}_m \mathbf{k}(\boldsymbol{\tau}, \mathbf{t}_m). \tag{II.2}$$

The sampling error and generalization error were computed to be

$$\begin{aligned} \text{sample error} &= 0.0724 \\ \text{generalization error} &= 0.4054 \end{aligned}$$

This value of $\sigma = 0.1$ is good because it allows the estimated function to pass through all the data points and at the same time preserve a similar trajectory to the true function. This is verified from the fact that both the sample error and generalization error are low enough to be a somewhat an optimal solution to the minimization problem.

Question (b)

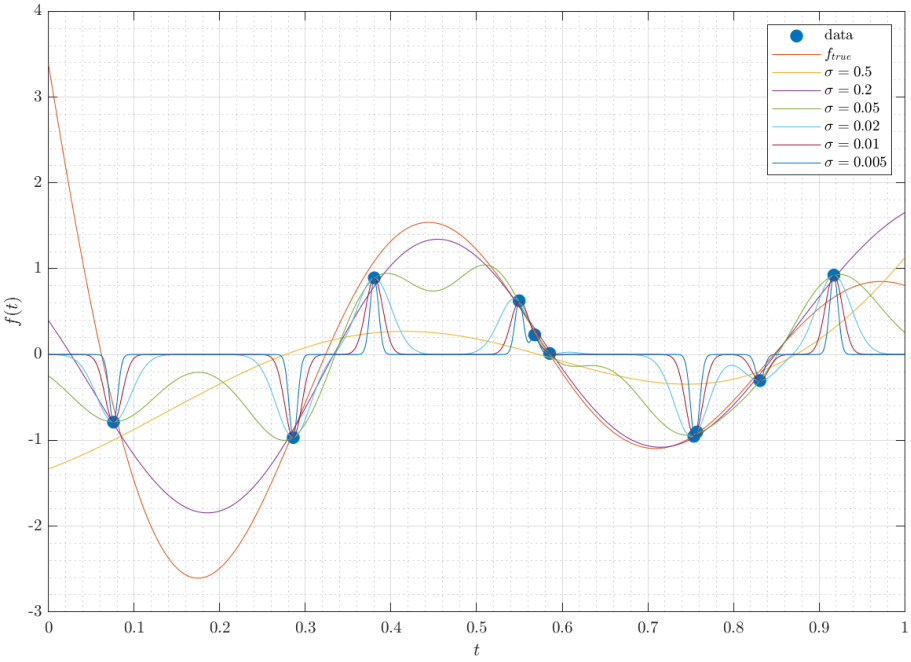

Figure 5: Plots of estimated $\hat{f}$ and f_{true} for multiple σ values.

Table 1: Sample and generalization errors.

σ	sample error	generalization error
0.5	1.7027	0.89238
0.2	0.1978	0.31412
0.05	0.047857	0.62799
0.02	0.011519	0.86954
0.01	0.0084172	0.95161
0.005	0.0082755	1

From Figure 5 and Table 1, we can tell the smaller the σ value is the more overfitting occurs for the estimate function. This explains how $\sigma = 0.1$ was optimal for our case where the number of data points is not as many. Hence, if the number of data points are much larger we can perhaps use a σ value that is larger to avoid underfitting.

III Problem Three

Consider the set of bump basis vectors $\psi_1(t), \ldots, \psi_N(t)$, where

$$\psi_k(t) = g\left(t - k/N\right), \quad g(t) = e^{-200t^2} \tag{III.1}$$

Given a point t, define the nonlinear "feature map" as

$$\mathbf{\Psi}(t) = \begin{bmatrix} \psi_1(t) \\ \psi_2(t) \\ \vdots \\ \psi_N(t) \end{bmatrix}$$

Plot the feature map as a discrete set of coefficients (In MATLAB, use `plot(1:N,Psit(1:N),'o')`). for $t = 1/3$ for $N = 10, 20, 50, 100, 200$. Compare to the radial basis kernel map

$$\mathbf{\Phi}_t(s) = k(s,t) = e^{-100|s-t|^2},$$

for $t = 1/3$ and $s \in [0,1]$. Discuss the relationship between kernel regression with a Gaussian radial basis function, and nonlinear regression using a basis of the form (III.1).

Solution

Question (a)

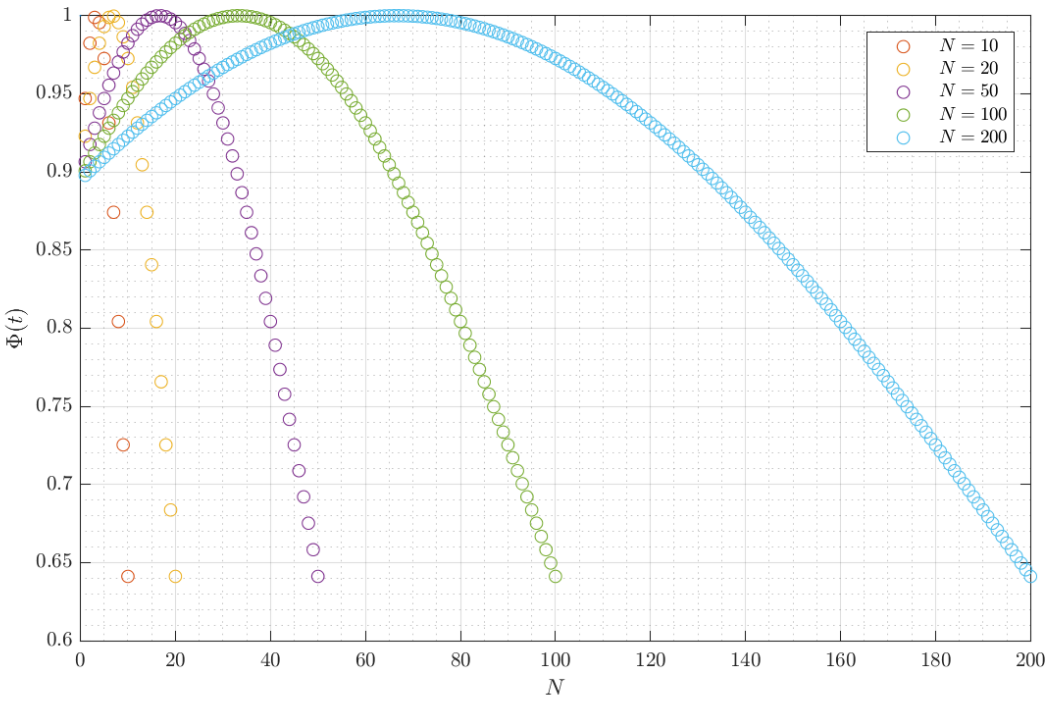

Figure 6: Feature map as a discrete set of coefficients.

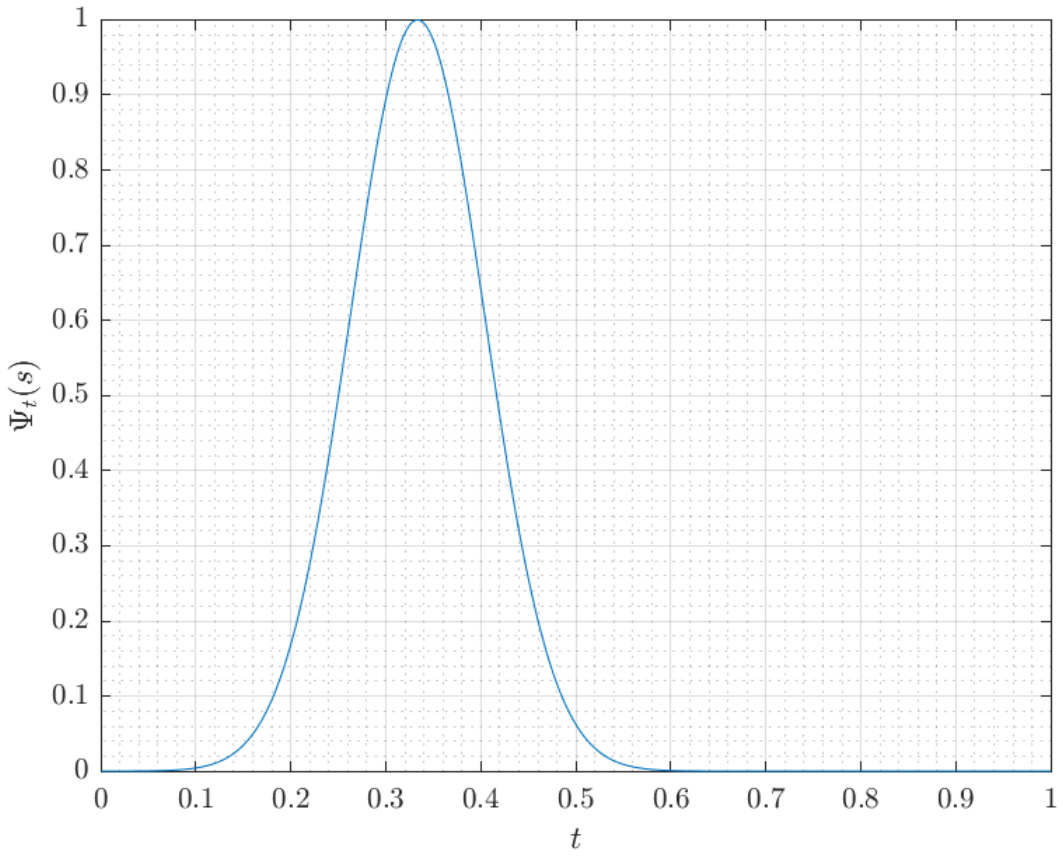

Figure 7: Plot of radial basis kernel map for $t = 1/3$ and $s \in [0, 1]$.

The bump function and the Gaussian radial basis function are very similar. However, the slight but crucial difference is that the bump function is for a finite-dimensional basis regression and the radial basis function is for a infinite-dimensional Hilbert space regression. The bump function solves for the regression problem by using it as bases and mapping the data points into the desired space. Whereas the radial basis function is a infinite dimensional mapping which is reduced to a finite linear combinations of this to approximate the optimal outputs from the data points by forming the reproducing kernel.

IV Problem Four

Let

$$\mathbf{A} = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 0.98 \end{bmatrix}$$

(a) Find the eigenvalue decomposition of $\mathbf{A}$ by hand. Recall that λ is an eigenvalue of $\mathbf{A}$ if for some $u[1], u[2]$ (entries of the corresponding eigenvector) we have

$$(1.01 - \lambda)u[1] + 0.99u[2] = 0$$
$$.99u[1] + (0.98 - \lambda)u[2] = 0.$$

Another way of saying this is that we want the values of λ such that $\mathbf{A} - \lambda\mathbf{I}$ (where $\mathbf{I}$ is the 2×2 identity matrix) has a non-trivial null space — there is a nonzero vector $\mathbf{u}$ such that $(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0$. Yet another way of saying this is that we want the values of λ such that $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. Once you have found the two eigenvalues, you can solve the 2×2 systems of equations $\mathbf{A}\mathbf{u}_1 = \lambda_1\mathbf{u}_1$ and $\mathbf{A}\mathbf{u}_2 = \lambda_2\mathbf{u}_2$ for $\mathbf{u}_1$ and $\mathbf{u}_2$.

Show your work above, but feel free to check you answer using MATLAB/numpy.

(b) If $\mathbf{y} = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, determine the solution to $\mathbf{A}\mathbf{x} = \mathbf{y}$.

(c) Now let $y = \begin{bmatrix} 1.1 & 1 \end{bmatrix}^\top$ and solve $\mathbf{A}\mathbf{x} = \mathbf{y}$. Comment on how the solution changed.

(d) Suppose we observe

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$$

with $\|\mathbf{e}\|_2 = 1$. We form an estimate $\tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{y}$. Which vector $\mathbf{e}$ (over all error vectors with $\|\mathbf{e}\|_2 = 1$) yields the maximum error $\|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2$?

(e) Which (unit) vector $\mathbf{e}$ yields the minimum error?

(f) Suppose the components of $\mathbf{e}$ are independent and identically distributed (i.i.d.) Gaussian random variables:

$$\mathbf{e} \sim \text{Normal}(0, \mathbf{I}).$$

What is the mean-square error $\mathbb{E}[\|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2]$?

(g) Verify your answer to part (f) in MATLAB/Python by taking $\mathbf{A}\mathbf{x} = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, and then generating $10,000$ different realizations of $\mathbf{e}$ using the `randn` command, and then averaging the results. Turn in your code and the results of your computation.

Solution

Question (a)

From $(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0$ we solve the characteristic polynomial obtained by $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$, that is

$$(1.01 - \lambda)(0.98 - \lambda) - 0.99^2 = 0$$
$$\lambda^2 - 1.99\lambda + 0.0097 = 0.$$

This gives us

$$\lambda_1 = 1.98511, \qquad \lambda_2 = 4.88637\text{e}{-3}.$$

Now we solve for $\text{Null}(\mathbf{A} - \lambda\mathbf{I})$ as follows.

$$\begin{bmatrix} 1.01 - \lambda_1 & 0.99 \\ 0.99 & 0.98 - \lambda_1 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0$$
$$\begin{bmatrix} 1 & -1.01528 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_{11} \\ v_{12} \end{bmatrix} = 0$$

Hence, we have

$$v_{11} - 1.01528 v_{12} = 0$$
$$v_{12} = 0$$

Then we have the following possible values

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 0.9850 \end{bmatrix}.$$

With the exact same approach we find

$$\begin{bmatrix} 1.01 - \lambda_2 & 0.99 \\ 0.99 & 0.98 - \lambda_2 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$
$$\begin{bmatrix} 1 & 0.99486 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} v_{21} \\ v_{22} \end{bmatrix} = 0$$

Hence, we have

$$v_{21} + 0.99486 v_{22} = 0$$
$$v_{22} = 0$$

Then we have the following possible values

$$\mathbf{v}_2 = \begin{bmatrix} -0.9949 \\ 1 \end{bmatrix}.$$

Thus, the eigenvalue decomposition becomes

$$\mathbf{A} = \begin{bmatrix} 1 & -0.9949 \\ 0.9850 & 1 \end{bmatrix} \begin{bmatrix} 1.9851 & 0 \\ 0 & 4.8864\text{e}{-3} \end{bmatrix} \begin{bmatrix} 1 & -0.9949 \\ 0.9850 & 1 \end{bmatrix}^{-1} \tag{IV.1}$$
$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{-1} \tag{IV.2}$$

Question (b)

Let $\{v_i\}_{i=1}^2$ be the column vectors of $\mathbf{V}$ or the eigenvector corresponding to each diagonal entry of the eigenvalue matrix, $\mathbf{\Lambda}$. From Sylvester's matrix theorem we have that

$$\mathbf{x} = \sum_{i=1}^{2} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^\top \mathbf{y} = \begin{bmatrix} -1.0309 \\ 2.0619 \end{bmatrix}.$$

Question (c)

With the exact same method in the previous problem we find that

$$\mathbf{x} = \sum_{i=1}^{2} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^\top \mathbf{y} = \begin{bmatrix} 9.0723 \\ -8.1443 \end{bmatrix}.$$

If we let the noise be a vector of $\mathbf{e} = \begin{bmatrix}0.1 & 0\end{bmatrix}^\top$, we can say $\mathbf{y} = \mathbf{Ax} + \mathbf{e}$ where $\mathbf{x}$ is the solution from Question (b). Say the solution from Question (b) be $\mathbf{x}_b$ and the solution from Question (c) be $\mathbf{x}_c$. Then we can see that the change of solution can be evaluated as

$$\|\mathbf{x}_c - \mathbf{x}_b\|_2^2 = \left\|\mathbf{A^{-1}e}\right\|_2^2. \tag{IV.3}$$

Now,

$$\left\| \begin{bmatrix} 9.0723 \\ -8.1443 \end{bmatrix} - \begin{bmatrix} -1.0309 \\ 2.0619 \end{bmatrix} \right\|_2^2 = 206.2387$$

and

$$\left\|\mathbf{A^{-1}e}\right\|_2^2 = 206.2387.$$

Thus, we can view the change between Question (b) and (c) to be given by some error of $\mathbf{e}$ introduced to the system $\mathbf{y} = \mathbf{Ax}$.

Question (d)

The maximum error can be found from (IV.3), which is

$$\left\|\mathbf{A}^{-1}\mathbf{e}\right\|_2^2 \leq \lambda_2^{-2} \|\mathbf{e}\|_2^2. \tag{IV.4}$$

If we let $\mathbf{A}^{-1} = \mathbf{B} \in \mathbb{R}^{2\times 2}$. Then we can find the noise vector corresponding to the maximum error by computing

$$(B_{11}e_1 + B_{12}e_2)^2 + (B_{21}e_1 + B_{22}e_2)^2 = \frac{1}{\lambda_2^2}(e_1^2 + e_2^2)$$
$$e_1^2 + e_2^2 = 1$$

This yields the solution of

$$\mathbf{e} = \begin{bmatrix} \mp 0.7017 \\ \pm 0.7124 \end{bmatrix} \tag{IV.5}$$

Question (e)

In contrast with the maximum the minimum is found by solving

$$\left\|\mathbf{A}^{-1}\mathbf{e}\right\|_2^2 \geq \lambda_1^{-2} \|\mathbf{e}\|_2^2. \tag{IV.6}$$

Thus,

$$(B_{11}e_1 + B_{12}e_2)^2 + (B_{21}e_1 + B_{22}e_2)^2 = \frac{1}{\lambda_1^2}(e_1^2 + e_2^2)$$
$$e_1^2 + e_2^2 = 1$$

This yields the solution of

$$\mathbf{e} = \begin{bmatrix} \pm 0.7124 \\ \pm 0.7017 \end{bmatrix} \tag{IV.7}$$

Question (f)

We know that the mean-square error can be found as

$$\mathbb{E}\left[\|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2\right] = \sigma^2 \sum_{i=1}^{2} \lambda_i^{-2} = \sum_{i=1}^{2} \lambda_i^{-2},$$

Since $\mathbf{e}$ is a standard normal variable with variance of 1. Hence we have

$$\mathbb{E}\left[\|\tilde{\mathbf{x}} - \mathbf{x}\|_2^2\right] = 41882.2404.$$

Question (g)

From a Monte Carlo test, we produce the following plot

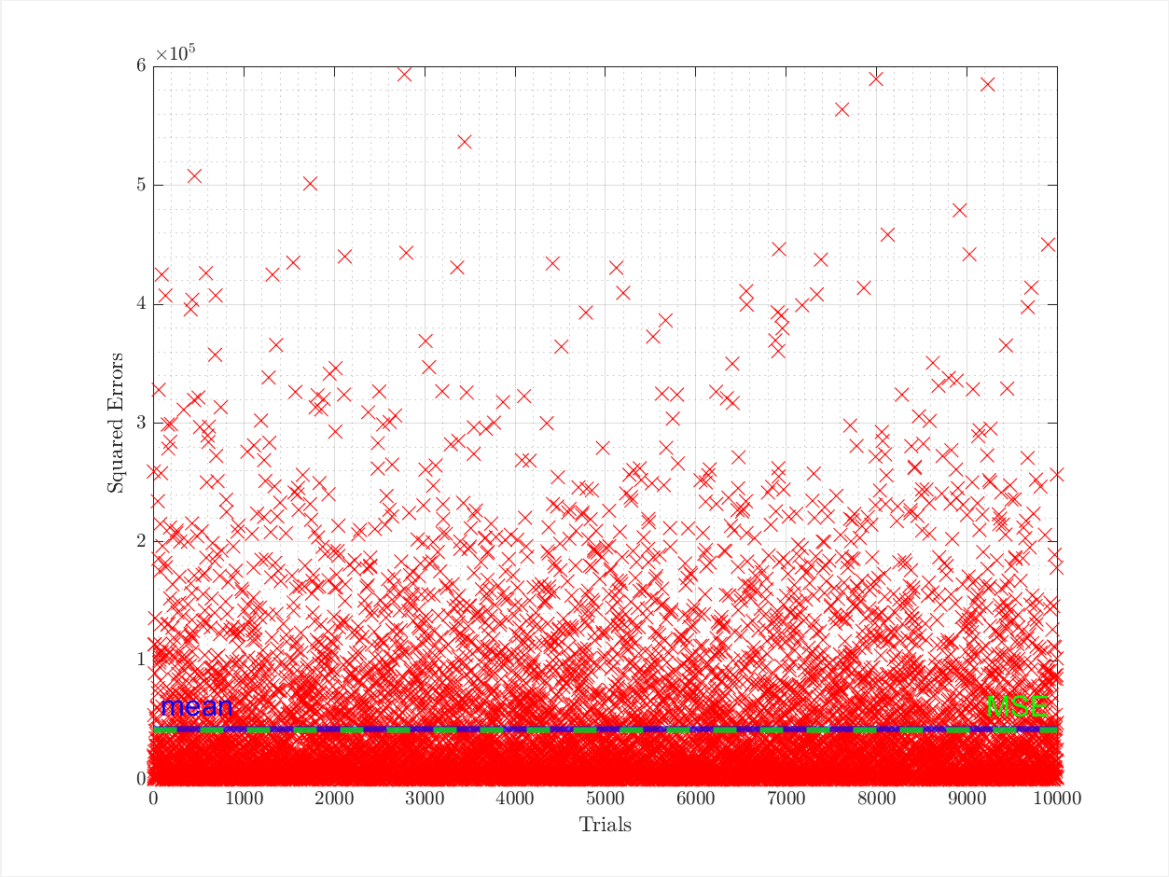

Figure 8: Monte Carlo test plot of squared error and its mean (blue) and the theoretical MSE value (green).

V Problem Five

(a) Let $\mathbf{A}$ be a $N \times N$ symmetric matrix. Show that (The trace of a (square) matrix is the sum of the elements on the diagonal: $\text{trace}(\mathbf{A}) = \sum_{n=1}^{N} A[n,n]$.)

$$\text{trace}(\mathbf{A}) = \sum_{n=1}^{N} \lambda_n,$$

where the $\{\lambda_n\}$ are the eigenvalues of $\mathbf{A}$.

(b) Now let $\mathbf{A}$ be an arbitrary $M \times N$ matrix. Recall the definition of the Frobenius norm:

$$\|\mathbf{A}\|_F = \left(\sum_{m=1}^{M} \sum_{n=1}^{N} |A[m,n]|^2 \right)^{1/2}.$$

Show that

$$\|\mathbf{A}\|_F^2 = \text{trace}(\mathbf{A}^\top \mathbf{A}) = \sum_{r=1}^{R} \sigma_r^2,$$

where R is the rank of $\mathbf{A}$ and the $\{\sigma_r\}$ are the singular values of $\mathbf{A}$.

(c) The *operator norm* (sometimes called the *spectral norm*) of an $M \times N$ matrix is

$$\|\mathbf{A}\| = \max_{\mathbf{x} \in \mathbb{R}^N,\ \|\mathbf{x}\|_2 = 1} \|\mathbf{A}\mathbf{x}\|_2.$$

(This matrix norm is so important, it doesn't even require a designation in its notation — if somebody says "matrix norm" and doesn't elaborate, this is what they mean.) Show that

$$\|\mathbf{A}\| = \sigma_1,$$

where σ_1 is the largest singular value of $\mathbf{A}$. For which $\mathbf{x}$ does

$$\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{A}\| \cdot \|\mathbf{x}\|_2 \quad ?$$

(d) Prove that $\|\mathbf{A}\| \leq \|\mathbf{A}\|_F$. Give an example of an $\mathbf{A}$ with $\|\mathbf{A}\| = \|\mathbf{A}\|_F$.

Solution

Question (a)

A symmetric matrix can be decomposed using its eigenvalues and eigenvectors as follows.

$$\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{H}}.$$

Now using the property of trace being invariant to cyclic permutation, $\text{tr}(\mathbf{ABC}) = \text{tr}(CAB) = \text{tr}(BCA)$, we have

$$\text{tr}\left(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathrm{H}}\right) = \text{tr}\left(\mathbf{V}^{\mathrm{H}}\mathbf{V}\mathbf{\Lambda}\right) = \text{tr}\left(\mathbf{V}^{-1}\mathbf{V}\mathbf{\Lambda}\right) = \text{tr}\left(\mathbf{\Lambda}\right) = \sum_{n=1}^{N} \lambda_n.$$

■

Question (b)

First we use the singular value decomposition as

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \tag{V.1}$$

where $\mathbf{U} \in \mathbb{R}^{M\times R}$, $\mathbf{V} \in \mathbb{R}^{N\times R}$, and $\mathbf{\Sigma} \in \mathbb{R}^{R\times R}$. Now

$$\begin{aligned}\mathbf{AV} &= \mathbf{U\Sigma}\\ \|\mathbf{AV}\|_F^2 &= \|\mathbf{U\Sigma}\|_F^2\\ \|\mathbf{A}\|_F^2 &= \|\mathbf{\Sigma}\|_F^2 = \sum_{r=1}^{R}\sigma_r^2.\end{aligned}$$

Next,

$$\begin{aligned}\mathrm{tr}\left(\mathbf{A}^\top\mathbf{A}\right) &= \mathrm{tr}\left((\mathbf{U\Sigma V}^\top)^\top\mathbf{U\Sigma V}^\top\right)\\ &= \mathrm{tr}\left(\mathbf{V\Sigma U}^\top\mathbf{U\Sigma V}^\top\right) = \mathrm{tr}\left(\mathbf{V\Sigma^2 V}^\top\right)\\ &= \mathrm{tr}\left(\mathbf{V}^\top\mathbf{V\Sigma^2}\right) = \mathrm{tr}\left(\mathbf{\Sigma}^2\right)\\ &= \sum_{r=1}^{R}\sigma_r^2.\end{aligned}$$

■

Question (c)

For this we have

$$\|\mathbf{A}\|^2 = \max_{\|\mathbf{x}\|_2=1}\frac{\|\mathbf{Ax}\|_2^2}{\|\mathbf{x}\|_2^2} = \max_{\|\mathbf{x}\|_2=1}\frac{\mathbf{x}^\top\mathbf{A}^\top\mathbf{Ax}}{\mathbf{x}^\top\mathbf{x}} = \lambda_1(\mathbf{A}^\top\mathbf{A}) = \sigma_1(\mathbf{A})^2$$ ■

Note that $\mathbf{U}$ is a unitary matrix for the singular value decomposition. Then

$$\max_{\|\mathbf{x}\|_2=1}\|\mathbf{Ax}\|_2 = \max_{\|\mathbf{x}\|_2=1}\left\|\mathbf{U\Sigma V}^\top\mathbf{x}\right\|_2 = \max_{\|\mathbf{x}\|_2=1}\left\|\mathbf{\Sigma V}^\top\mathbf{x}\right\|_2.$$

Now let $\mathbf{y} = \mathbf{V}^\top\mathbf{x}$. By the same argument that the unitary matrix $\|\mathbf{Uz}\| = \mathbf{z}^\top\mathbf{U}^\top\mathbf{Uz} = \mathbf{z}^\top\mathbf{z} = \|z\|_2^2$ for $\mathbf{z} \in \mathbb{R}^N$, we have $\|\mathbf{y}\|_2 = \left\|\mathbf{V}^\top\mathbf{x}\right\|_2 = \|\mathbf{x}\|_2 = 1$ since $\mathbf{V}$ is unitary. Thus,

$$\max_{\|\mathbf{x}\|_2=1}\left\|\mathbf{\Sigma V}^\top\mathbf{x}\right\|_2 = \max_{\|\mathbf{y}\|_2=1}\|\mathbf{\Sigma y}\|$$

Then since $\mathbf{\Sigma} = \mathrm{diag}\,(\sigma_1, ..., \sigma_n)$, where $\sigma_1 > \cdots > \sigma_n$. The maximum is attained when $\mathbf{y} = \begin{bmatrix}1 & \cdots & 0\end{bmatrix}^\top$ which gives σ_1. And since $\mathbf{y} = \mathbf{V}^\top\mathbf{x}$ that is when $\mathbf{x} = \mathbf{v}_1$ where $\mathbf{v}_1$ is the singular vector (column vector in $\mathbf{V}$) that corresponds to the singular value of σ_1. Hence, $\|\mathbf{Ax}\|_2 = \|\mathbf{A}\|\cdot\|\mathbf{x}\|_2$ is achieved when $\mathbf{x} = \mathbf{v}_1$.

Question (d)

In Question (b) we have shown that

$$\|\mathbf{A}\|_F = \sqrt{\sum_{r=1}^{R}\sigma_r^2}$$

and in the previous question we find that $\|\mathbf{A}\| = \sigma_1$. Hence, it naturally follows that

$$\|\mathbf{A}\| \le \|\mathbf{A}\|_F.$$

With the following code we can produce a $\mathbf{A} \in \mathbb{R}^{2\times 2}$ with singular values of [1,0]. Which is an example of $\|\mathbf{A}\| = 1 = \|\mathbf{A}\|_F$.

```
H = hilb(2);
V = orth(H);
L = [1 0; 0 0];
A = V * L * V.'
B = A.'*A;
[U,S,V] = svd(B);
```

$$\mathbf{A} = \begin{bmatrix} 0.7774 & 0.4160 \\ 0.4160 & 0.2226 \end{bmatrix} = \begin{bmatrix} -0.8817 & -0.4719 \\ -0.4719 & 0.8817 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} -0.8817 & -0.4719 \\ -0.4719 & 0.8817 \end{bmatrix}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$$