Introduction to Machine Learning
Fall 2022
University of Science and Technology of China

Lecturer: Jie Wang
Name: Yunqin Zhu

Homework 7
ID: PB20061372

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Singular Value Decomposition

Let $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\text{rank}\,(\mathbf{A}) = r$. The SVD of $\mathbf{A}$ is $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{m\times m}, \mathbf{\Sigma} \in \mathbb{R}^{m\times n}, \mathbf{V} \in \mathbb{R}^{n\times n}$, and we sort the diagonal entries of $\mathbf{\Sigma}$ in the descending order $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$. Denote

$$\mathbf{U_1} = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r),\ \mathbf{U_2} = (\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m),$$
$$\mathbf{V_1} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r),\ \mathbf{V_2} = (\mathbf{v}_{r+1}, \ldots, \mathbf{v}_n).$$

The column space of $\mathbf{A}$ is the set

$$\mathcal{C}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{A}\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\}.$$

The null space of $\mathbf{A}$ is the set

$$\mathcal{N}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^n : \mathbf{A}\mathbf{y} = \mathbf{0}\}.$$

1. Show that

   (a) $P_{\mathcal{C}(\mathbf{A})}(\mathbf{x}) = \mathbf{U_1}\mathbf{U_1}^\top\mathbf{x}$;

   (b) $P_{\mathcal{N}(\mathbf{A})}(\mathbf{x}) = \mathbf{V_2}\mathbf{V_2}^\top\mathbf{x}$;

   (c) $P_{\mathcal{C}(\mathbf{A}^\top)}(\mathbf{x}) = \mathbf{V_1}\mathbf{V_1}^\top\mathbf{x}$;

   (d) $P_{\mathcal{N}(\mathbf{A}^\top)}(\mathbf{x}) = \mathbf{U_2}\mathbf{U_2}^\top\mathbf{x}$.

   Solution:

   (a) $P_{\mathcal{C}(\mathbf{A})}(\mathbf{x}) = P_{\mathcal{C}(\mathbf{U_1})}(\mathbf{x}) = \mathbf{U_1}\left(\mathbf{U_1}^\top\mathbf{U_1}\right)^{-1}\mathbf{U_1}^\top\mathbf{x} = \mathbf{U_1}\mathbf{I_r}^{-1}\mathbf{U_1}^\top\mathbf{x} = \mathbf{U_1}\mathbf{U_1}^\top\mathbf{x}$.

   (b) $P_{\mathcal{N}(\mathbf{A})}(\mathbf{x}) = P_{\mathcal{C}(\mathbf{A}^\top)^\perp}(\mathbf{x}) = P_{\mathcal{C}(\mathbf{V_2})}(\mathbf{x}) = \mathbf{V_2}\left(\mathbf{V_2}^\top\mathbf{V_2}\right)^{-1}\mathbf{V_2}^\top\mathbf{x} = \mathbf{V_2}\mathbf{V_2}^\top\mathbf{x}$.

   (c) $P_{\mathcal{C}(\mathbf{A}^\top)}(\mathbf{x}) = P_{\mathcal{C}(\mathbf{V_1})}(\mathbf{x}) = \mathbf{V_1}\left(\mathbf{V_1}^\top\mathbf{V_1}\right)^{-1}\mathbf{V_1}^\top\mathbf{x} = \mathbf{V_1}\mathbf{I_r}^{-1}\mathbf{V_1}^\top\mathbf{x} = \mathbf{V_1}\mathbf{V_1}^\top\mathbf{x}$.

   (d) $P_{\mathcal{N}(\mathbf{A}^\top)}(\mathbf{x}) = P_{\mathcal{C}(\mathbf{A})^\perp}(\mathbf{x}) = P_{\mathcal{C}(\mathbf{U_2})}(\mathbf{x}) = \mathbf{U_2}\left(\mathbf{U_2}^\top\mathbf{U_2}\right)^{-1}\mathbf{U_2}^\top\mathbf{x} = \mathbf{U_2}\mathbf{U_2}^\top\mathbf{x}$. ■

2. The Frobenius norm of $\mathbf{A}$ is

$$\|\mathbf{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}^2}.$$

   (a) Show that $\|\mathbf{A}\|_F^2 = \text{tr}\left(\mathbf{A}^\top\mathbf{A}\right)$.

(b) Let $\mathbf{B} \in \mathbb{R}^{m\times n}$. Suppose that $\mathcal{C}(\mathbf{A})\perp\mathcal{C}(\mathbf{B})$, that is,

$$\langle \mathbf{a}, \mathbf{b}\rangle = 0,\ \forall\, \mathbf{a} \in \mathcal{C}(\mathbf{A}),\ \mathbf{b} \in \mathcal{C}(\mathbf{B}).$$

Show that

$$\|\mathbf{A}+\mathbf{B}\|_F^2 = \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2.$$

Solution:

(a)
$$\begin{aligned}\operatorname{tr}\left(\mathbf{A}^\top\mathbf{A}\right) &= \operatorname{tr}\begin{pmatrix} \sum_{i=1}^m a_{i,1}^2 & \sum_{i=1}^m a_{i,1}a_{i,2} & \cdots & \sum_{i=1}^m a_{i,1}a_{i,n} \\ \sum_{i=1}^m a_{i,1}a_{i,2} & \sum_{i=1}^m a_{i,2}^2 & \cdots & \sum_{i=1}^m a_{i,2}a_{i,n} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^m a_{i,1}a_{i,n} & \sum_{i=1}^m a_{i,2}a_{i,n} & \cdots & \sum_{i=1}^m a_{i,n}^2 \end{pmatrix} \\ &= \sum_{i=1}^m\sum_{j=1}^n a_{i,j}^2 = \|\mathbf{A}\|_F^2.\end{aligned}$$

(b)
$$\begin{aligned}\|\mathbf{A}+\mathbf{B}\|_F^2 &= \sum_{i=1}^n \|\mathbf{a_i}+\mathbf{b_i}\|^2 \\ &= \sum_{i=1}^n \left(\|\mathbf{a}_i\|^2 + \langle \mathbf{a_i}, \mathbf{b_i}\rangle + \|\mathbf{b}_i\|^2\right) \\ &= \sum_{i=1}^n \|\mathbf{a}_i\|^2 + \sum_{i=1}^n \|\mathbf{b}_i\|^2 \\ &= \|\mathbf{A}\|_F^2 + \|\mathbf{B}\|_F^2.\end{aligned}$$

■

3. Given $K < r$, $K \in \mathbb{N}$, please solve the problem as follows

$$\min_{\mathbf{X}\in\mathbb{R}^{m\times n}} \{\|\mathbf{A}-\mathbf{X}\|_F : \operatorname{rank}(\mathbf{X}) \le K\}.$$

For simplicity, you can assume that all singular values of $\mathbf{A}$ are different.

Solution:
Any $\mathbf{X} \in \mathbb{R}^{m\times n}$ with $\operatorname{rank}(\mathbf{X}) \le K$ can be decomposed as $\mathbf{X} = \mathbf{QR}$, where $\mathbf{Q} \in \mathbb{R}^{m\times K}$ has orthonormal columns and $\mathbf{R} \in \mathbb{R}^{K\times n}$. Then, the optimization problem can be rewritten as

$$\min_{\mathbf{Q}\in\mathbb{R}^{m\times K},\mathbf{R}\in\mathbb{R}^{K\times n}} \{\|\mathbf{A}-\mathbf{QR}\|_F^2 : \mathbf{Q}^\top\mathbf{Q} = \mathbf{I}\}.$$

Taking the derivative of $\|\mathbf{A}-\mathbf{QR}\|_F^2$ with respect to $\mathbf{R}$, we obtain a necessary optimality condition

$$\mathbf{Q}^\top\mathbf{A} - \mathbf{Q}^\top\mathbf{QR} = 0 \implies \mathbf{R} = \mathbf{Q}^\top\mathbf{A}.$$

Plugging it into the objective function, we have

$$\|\mathbf{A}-\mathbf{QR}\|_F^2 = \|\mathbf{A}-\mathbf{QQ}^\top\mathbf{A}\|_F^2 = \|\mathbf{A}\|_F^2 - \|\mathbf{Q}^\top\mathbf{A}\|_F^2,$$

where the first term is a constant and the second term is

$$-\|\mathbf{Q}^\top \mathbf{A}\|_F^2 = -\operatorname{tr}\left(\mathbf{Q}^\top \mathbf{A}\mathbf{A}^\top \mathbf{Q}\right) = -\operatorname{tr}\left(\mathbf{Q}^\top \mathbf{U}\mathbf{\Sigma}^2\mathbf{U}^\top \mathbf{Q}\right).$$

Denote $\mathbf{P} = \mathbf{U}^\top \mathbf{Q}$, and hence the optimization problem becomes

$$\max_{\mathbf{P}\in\mathbb{R}^{m\times K}} \left\{\operatorname{tr}\left(\mathbf{P}^\top \mathbf{\Sigma}^2 \mathbf{P}\right) : \mathbf{P}^\top \mathbf{P} = \mathbf{I}\right\},$$

which has been solved in Lecture 16. The optimal solution $\mathbf{P}^*$ has the form

$$\mathbf{P}^* = \begin{pmatrix} \tilde{\mathbf{P}}^* \\ \mathbf{O} \end{pmatrix},\ \forall \tilde{\mathbf{P}}^* \in \mathbb{R}^{K\times K},\ \tilde{\mathbf{P}}^{*\top}\tilde{\mathbf{P}}^* = \mathbf{I}.$$

Therefore, the optimal solution to the original problem is

$$\begin{aligned}\mathbf{X}^* &= \mathbf{Q}\mathbf{Q}^\top \mathbf{A} = \mathbf{U}\mathbf{P}^{*\top}\mathbf{P}^*\mathbf{U}^\top \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \\ &= \mathbf{U}\begin{pmatrix} \mathbf{I}_K & \\ & \mathbf{O} \end{pmatrix}\mathbf{\Sigma}\mathbf{V}^\top = \sum_{i=1}^{K} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top.\end{aligned}$$

The optimal value is

$$\|\mathbf{A} - \mathbf{X}^*\|_F = \sqrt{\left\|\sum_{i=K+1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top\right\|_F^2} = \sqrt{\sum_{i=K+1}^{r} \sigma_i^2 \operatorname{tr}\left(\mathbf{u}_i \mathbf{v}_i^\top \mathbf{v}_i \mathbf{u}_i^\top\right)} = \sqrt{\sum_{i=K+1}^{r} \sigma_i^2}. \qquad \blacksquare$$

4. **Programming Exercise.** We provide you with a color image ("Hinton.jpg"). Suppose that $\mathbf{A} = (\mathbf{A}_i)_{i=1}^3$ is the data tensor of the image, where $\mathbf{A}_i, i = 1, 2, 3$, represents different channels. We have each $\mathbf{A}_i \in \mathbb{R}^{500\times 500}$ and $r = \operatorname{rank}(\mathbf{A}_i) = 500$. In this exercise, you are expected to implement an image compression algorithm following the steps below. You can use your favorite programming language.

   (a) Compute the SVD $\mathbf{A}_i = \mathbf{U}_i\mathbf{\Sigma}_i\mathbf{V}_i^\top = \sum_{j=1}^{r} \sigma_{i,j}\mathbf{u}_{i,j}\mathbf{v}_{i,j}^\top$, where $i = 1, 2, 3$, $\sigma_{i,1} \geq \sigma_{i,2} \geq \cdots \geq \sigma_{i,r} > 0$ are the diagonal entries of $\mathbf{\Sigma}_i$, $\mathbf{u}_{i,j}$ is the jth column of $\mathbf{U}_i$, and $\mathbf{v}_{i,j}$ is the jth column of $\mathbf{V}_i$.

   (b) Use the first k ($k < r$) terms of SVD to approximate the original image $\mathbf{A}$. Then, we get the compressed images, the data tensors of which are $\mathbf{A}_k = (\mathbf{A}_{i,k})_{i=1}^3 = (\sum_{j=1}^{k} \sigma_{i,j}\mathbf{u}_{i,j}\mathbf{v}_{i,j}^\top)_{i=1}^3$. Compute $||\mathbf{A} - \mathbf{A}_k||_F$, i.e., $\sum_{i=1}^3 ||\mathbf{A}_i - \mathbf{A}_{i,k}||_F$, for $k = 2, 4, 8, 16, 32, 64, 128, 256$.

   (c) Plot $\mathbf{A}_k$ as images for all the ks in (b).

Solution:
A Python implementation is given in "HW7.ipynb". The computed singular values are shown in Figure 2 in descending order. The compressed images and the corresponding $\|\mathbf{A} - \mathbf{A}_k\|_F$ are shown in Figure 3. Note that we normalize all RGB values to be in the range $[0, 1]$. $\blacksquare$

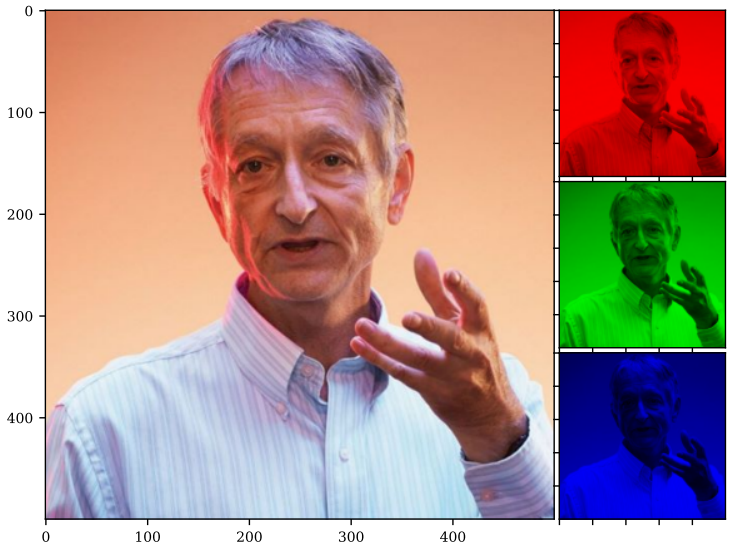

Figure 1: Hinton with RGB channels.

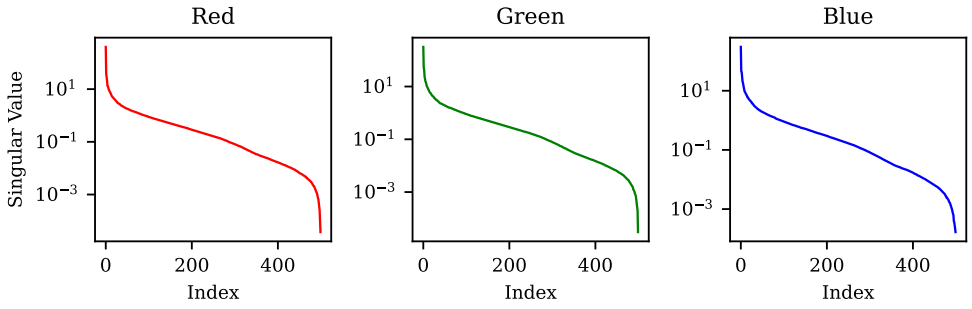

Figure 2: Singular values of the three channels.

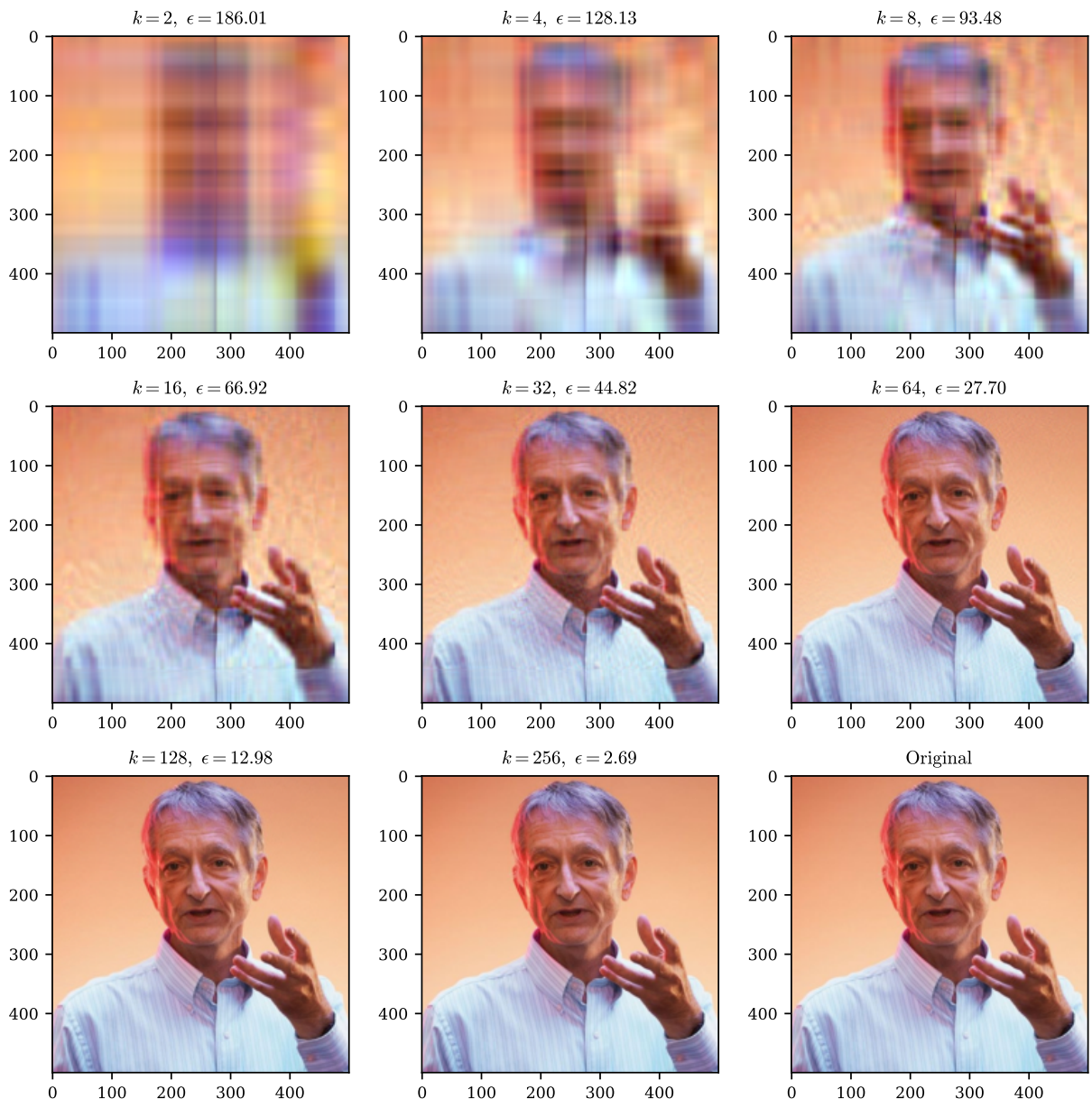

Figure 3: Compressed images and the corresponding truncation errors $\epsilon = \|\mathbf{A} - \mathbf{A}_k\|_F$.

Exercise 2: Principle Component Analysis

Suppose that we have a set of data instances $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$. Let $\widetilde{\mathbf{X}} \in \mathbb{R}^{d\times n}$ be the matrix whose i^{th} column is $\mathbf{x}_i - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ is the sample mean, and $\mathbf{S}$ be the sample variance matrix.

1. For $\mathbf{G} \in \mathbb{R}^{d\times K}$, let us define

$$f(\mathbf{G}) = \operatorname{tr}\left(\mathbf{G}^\top \mathbf{S}\mathbf{G}\right). \tag{1}$$

Show that $f(\mathbf{GQ}) = f(\mathbf{G})$ for any orthogonal matrix $\mathbf{Q} \in \mathbb{R}^{K\times K}$.

Solution:
$f(\mathbf{GQ}) = \operatorname{tr}\left(\mathbf{Q}^\top \mathbf{G}^\top \mathbf{S}\mathbf{G}\mathbf{Q}\right) = \operatorname{tr}\left(\mathbf{Q}\mathbf{Q}^\top \mathbf{G}^\top \mathbf{S}\mathbf{G}\right) = \operatorname{tr}\left(\mathbf{G}^\top \mathbf{S}\mathbf{G}\right) = f(\mathbf{G})$, where we use the fact that $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$. ■

2. Please find $\mathbf{g}_1$ defined as follows by the Lagrange multiplier method.

$$\mathbf{g}_1 := \underset{\mathbf{g}\in\mathbb{R}^d}{\mathbf{argmax}} \{f(\mathbf{g}) : \|\mathbf{g}\|_2 = 1\}, \tag{2}$$

where f is defined by (1). Notice that, the vector $\mathbf{g}_1$ is the first principal component vector of the data.

Solution:
The Lagrangian of (2) is

$$L(\mathbf{g}, \lambda) = \mathbf{g}^\top \mathbf{S}\mathbf{g} - \lambda\left(\mathbf{g}^\top \mathbf{g} - 1\right).$$

Taking the derivative with respect to $\mathbf{g}$, we have

$$\nabla_{\mathbf{g}} L(\mathbf{g}, \lambda) = 2\mathbf{S}\mathbf{g} - 2\lambda\mathbf{g} = 0,$$

which gives $\mathbf{S}\mathbf{g} = \lambda\mathbf{g}$, a necessary condition for Lagrangian optimality. Plugging it into $L(\mathbf{g}, \lambda)$, we obtain the dual function

$$q(\lambda) = \max_{\mathbf{g}\in\mathbb{R}^d} \left\{\mathbf{g}^\top \mathbf{S}\mathbf{g} - \lambda\left(\mathbf{g}^\top \mathbf{g} - 1\right)\right\} = \max_{\mathbf{g}\in\mathbb{R}^d} \lambda = \lambda,$$
$$\mathbf{dom}\ q \subset \{\lambda \in \mathbb{R} : \lambda \text{ is an eigenvalue of } \mathbf{S}\}.$$

Note that, to maximize $L(\mathbf{g}, \lambda)$ with respect to $\mathbf{g}$, we must have

$$\nabla^2_{\mathbf{g}} L(\mathbf{g}, \lambda) = 2(\mathbf{S} - \lambda\mathbf{I}) \leq 0.$$

So $\lambda \geq \lambda_1(\mathbf{S})$. That is, $\mathbf{dom}\ q = \{\lambda_1(\mathbf{S})\}$, and hence the dual optimal value $\min_\lambda q(\lambda) = \lambda_1(\mathbf{S})$. Moreover, when $\mathbf{g}$ is a unit eigenvector corresponding to $\lambda_1(\mathbf{S})$, the primal objective $f(\mathbf{g}) = \lambda_1(\mathbf{S})\mathbf{g}^\top\mathbf{g} = \lambda_1(\mathbf{S}) = \min_\lambda q(\lambda)$, from which we conclude that the primal optimum is also achieved and there is no duality gap.

To sum up, $\mathbf{g}_1$ is a unit eigenvector corresponding to $\lambda_1(\mathbf{S})$, or equivalently, a left singular vector of $\widetilde{\mathbf{X}}$ corresponding to $\sigma_1(\widetilde{\mathbf{X}})$. ■

3. Please find $\mathbf{g}_2$ defined as follows by the Lagrange multiplier method.

$$\mathbf{g}_2 := \underset{\mathbf{g}\in\mathbb{R}^d}{\mathbf{argmax}}\,\{f(\mathbf{g}) : \|\mathbf{g}\|_2 = 1, \langle \mathbf{g}, \mathbf{g}_1\rangle = 0\}, \tag{3}$$

where $\mathbf{g}_1$ is given by (2). Similar to $\mathbf{g}_1$, the vector $\mathbf{g}_2$ is the second principal component vector of the data.

Solution:
Consider the spectral decomposition $\mathbf{S} = \sum_{i=1}^d \lambda_i \mathbf{g}_i \mathbf{g}_i^\top$, where $\mathbf{g}_i$ is the i-th largest eigenvector of $\mathbf{S}$ and λ_i is the corresponding eigenvalue. Hence, the objective of (3) becomes

$$f(\mathbf{g}) = \sum_{i=1}^d \lambda_i \langle \mathbf{g}, \mathbf{g}_i\rangle^2 = \sum_{i=2}^d \lambda_i \langle \mathbf{g}, \mathbf{g}_i\rangle^2 = \mathbf{g}^\top(\mathbf{S} - \mathbf{S}_1)\mathbf{g}.$$

By the same approach as used in (2), we can solve the following problem

$$\max_{\mathbf{g}\in\mathbb{R}^d}\{\mathbf{g}^\top(\mathbf{S} - \mathbf{S}_1)\mathbf{g} : \|\mathbf{g}\|_2 = 1\},$$

whose optimal solution $\mathbf{g}_2$ is a unit eigenvector corresponding to the largest eigenvalue of $\mathbf{S} - \mathbf{S}_1$, i.e. $\lambda_2(\mathbf{S})$, or equivalently, a left singular vector of $\widetilde{\mathbf{X}}$ corresponding to $\sigma_2(\widetilde{\mathbf{X}})$. Since $\mathbf{g}_2$ is in the feasible set of (3), it is also the optimal solution of (3). ■

4. Please derive the first K principal component vectors by repeating the above process.

Solution:
Given the following optimization problem

$$\mathbf{g}_k = \underset{\mathbf{g}\in\mathbb{R}^d}{\mathbf{argmax}}\,\{f(\mathbf{g}) : \|\mathbf{g}\|_2 = 1, \langle \mathbf{g}, \mathbf{g}_1\rangle = \cdots = \langle \mathbf{g}, \mathbf{g}_{k-1}\rangle = 0\}, \tag{4}$$

where $\mathbf{g}_1, \ldots, \mathbf{g}_{k-1}$ are the first $k-1$ principal component vectors, we reduce the objective to

$$f(\mathbf{g}) = \sum_{i=k}^d \lambda_i \langle \mathbf{g}, \mathbf{g}_i\rangle^2 = \mathbf{g}^\top(\mathbf{S} - \mathbf{S}_{k-1})\mathbf{g}.$$

We then relax the constraint and solve the following problem using the approach in (2)

$$\max_{\mathbf{g}\in\mathbb{R}^d}\{\mathbf{g}^\top(\mathbf{S} - \mathbf{S}_{k-1})\mathbf{g} : \|\mathbf{g}\|_2 = 1\},$$

whose optimal solution $\mathbf{g}_k$ is a unit eigenvector corresponding to the largest eigenvalue of $\mathbf{S} - \mathbf{S}_{k-1}$, i.e. $\lambda_k(\mathbf{S})$, or equivalently, a left singular vector of $\widetilde{\mathbf{X}}$ corresponding to $\sigma_k(\widetilde{\mathbf{X}})$. Since $\mathbf{g}_k$ is in the feasible set of (4), it is also the optimal solution of (4). By induction on $k = 2, \ldots, K$, we find the first K principal component vectors. ■

5. What is $f(\mathbf{g}_k)$, $k = 1, \ldots, K$? What about their meaning?

Solution:
$f(\mathbf{g}_k) = \mathbf{g}^\top\mathbf{S}\mathbf{g} = \lambda_k\mathbf{g}^\top\mathbf{g} = \lambda_k$, which is the k-th largest eigenvalue of $\mathbf{S}$, or equivalently, the square of the k-th largest singular value of $\frac{1}{\sqrt{n-1}}\widetilde{\mathbf{X}}$. ■

6. When are the first K principal component vectors unique?

   Solution:
   The first K principal component vectors, i.e. the unit eigenvectors of $\mathbf{S}$ corresponding to the K largest eigenvalues, are unique up to multiplication by -1, if and only if the K largest eigenvalues of $\mathbf{S}$ are distinct, and different from the remaining eigenvalues.

   Equivalently, the K largest singular values of $\widetilde{\mathbf{X}}$ are distinct and different from the remaining singular values.

   In such cases, the eigenspace corresponding to each of these eigenvalues is one-dimensional and has a unique basis up to multiplication by -1.

   Otherwise, there are at least two eigenvectors corresponding to the same eigenvalue, and any linear combination of these eigenvectors is also a principal component vector. ■

7. **Programming Exercise.** We provide you with 130 handwritten 3s, each a digitized 28×28 grayscale image("imgs_3"). Please finish the following steps. You can use your favorite programming language.

   (a) Consider these images as points $\mathbf{x}_i \in \mathbb{R}^{784}, i = 1, \cdots, 130$. Let $\bar{\mathbf{x}}$ be the mean of all $\mathbf{x}_i$. Let $\mathbf{X} \in \mathbb{R}^{130\times 784}$ with $\mathbf{x}_i - \bar{\mathbf{x}}$ as its i^{th} row.

   (b) Calculate the SVD, $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Let $\mathbf{v}_1, \mathbf{v}_2$ be the columns of $\mathbf{V}$ corresponding to the 2 largest singular values, respectively. Please show $\bar{\mathbf{x}}, \mathbf{v}_1$ and $\mathbf{v}_2$, considering them as 28×28 grayscale images.

   (c) What do the three images illustrate?

   Solution:
   A Python implementation is given in "HW7.ipynb". The computed singular values are shown in Figure 5 in descending order. The image of $\bar{\mathbf{x}}, \mathbf{v}_1$ and $\mathbf{v}_2$ are shown in Figure 6. Note that we normalize all grayscale values to be in the range $[0, 1]$ and show 1 as black and 0 as white.

   The image of $\bar{\mathbf{x}}$, the mean of data vectors, illustrates the average of all 130 images.

   The image of $\mathbf{v}_1$, the first pricipal component vector, illustrates the direction of the largest variation of the 130 images.

   The image of $\mathbf{v}_2$, the second pricipal component vector, illustrates the direction of the second largest variation of the 130 images, which is orthogonal to that of $\mathbf{v}_1$. ■

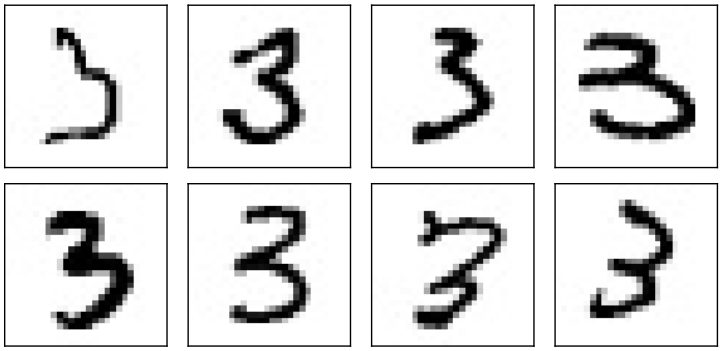

Figure 4: Some samples of the images.

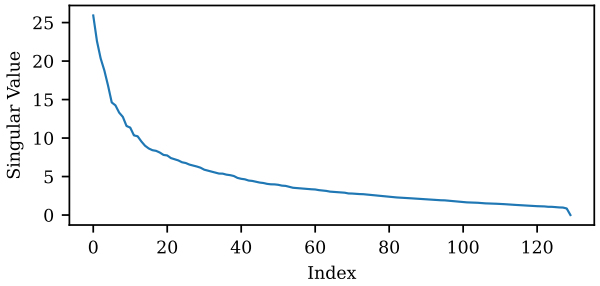

Figure 5: Singular values of $\mathbf{X}$.

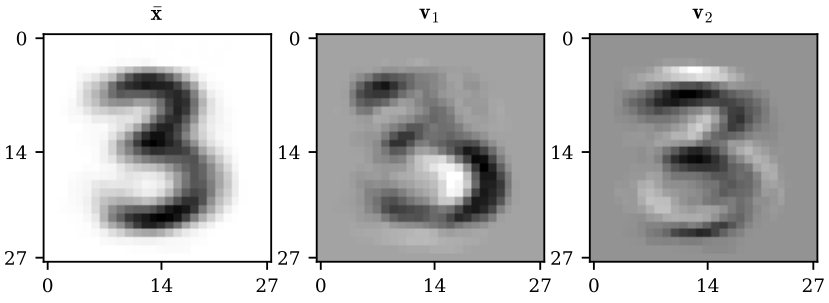

Figure 6: Mean and the first 2 principal components of $\mathbf{X}$.

Exercise 3: Properties of Transition Matrix

A transition matrix (also called a stochastic matrix, probability matrix) is a square matrix used to describe the transitions of a Markov chain. Each of its entries is a non-negative real number representing a probability. A right (left) transition matrix is a square matrix with each row (column) summing to one. Without loss of generality, we study the right transition matrix in this exercise. Suppose that $\mathbf{T} \in \mathbb{R}^{n\times n}$ is a right transition matrix.

1. Show that 1 is an eigenvalue of $\mathbf{T}$.

   Solution:
   Since the sum of each row equals to 1, we have $\mathbf{T1} = \mathbf{1}$, i.e. 1 is an eigenvalue of $\mathbf{T}$ with the corresponding eigenvector being $\mathbf{1}$. ■

2. Let λ be an eigenvalue of $\mathbf{T}$. Show that $|\lambda| \leq 1$.

   Solution:
   Let λ and $\mathbf{x}$ be a pair of eigenvalue and eigenvector of $\mathbf{T}$. Then $\mathbf{Tx} = \lambda\mathbf{x}$. Taking the infinity norm of both sides, we have

   $$|\lambda|\|\mathbf{x}\|_\infty = \|\mathbf{Tx}\|_\infty \leq \|\mathbf{T}\|_\infty\|\mathbf{x}\|_\infty = \|\mathbf{x}\|_\infty,$$

   where $\|\mathbf{T}\|_\infty = 1$ as each row of $\mathbf{T}$ sums to 1. Therefore, $|\lambda| \leq 1$. ■

3. Show that $\mathbf{I} - \gamma\mathbf{T}$ is invertible, where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix and $\gamma \in (0, 1)$.

   Solution:
   Suppose that λ is an eigenvalue of $\mathbf{T}$ and $\mathbf{x}$ is the corresponding eigenvector. Since $\mathbf{Tx} = \lambda\mathbf{x}$ is equivalent to $(\mathbf{I} - \gamma\mathbf{T})\mathbf{x} = (1 - \gamma\lambda)\mathbf{x}$, we conclude that the eigenvalue of $\mathbf{I} - \gamma\mathbf{T}$ must has the form of $1 - \gamma\lambda$. Note that $|1 - \gamma\lambda| \geq 1 - \gamma|\lambda| \in (0, 1)$. That is, any eigenvalue $1 - \gamma\lambda$ is non-zero. Therefore, $\mathbf{I} - \gamma\mathbf{T}$ is invertible. ■

4. We will show that $(\mathbf{I} - \gamma\mathbf{T})^{-1} = \sum_{i=0}^{\infty}(\gamma\mathbf{T})^i$.

   (a) For $\mathbf{x} \in \mathbb{R}^n$, the infinity norm is defined by

   $$\|\mathbf{x}\|_\infty = \max_i |x_i|.$$

   The induced norm of a matrix $\mathbf{M} \in \mathbb{R}^{m\times n}$ is

   $$\|\mathbf{M}\|_\infty = \max_{\|\mathbf{x}\|_\infty \leq 1} \|\mathbf{Mx}\|_\infty.$$

   i. Show that $\|\mathbf{M}\|_\infty = \max_i \sum_{j=1}^{n} |m_{i,j}|$.

   ii. Show that $\|\mathbf{AB}\|_\infty \leq \|\mathbf{A}\|_\infty\|\mathbf{B}\|_\infty$ holds for any $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{B} \in \mathbb{R}^{n\times p}$.

   (b) Show that the sequence $\left\{\sum_{i=0}^{k}(\gamma\mathbf{T})^i\right\}_{k=0}^{\infty}$ converges.

   (c) Let $\left\{\sum_{i=0}^{k}(\gamma\mathbf{T})^i\right\}_{k=0}^{\infty}$ converges to a matrix L. Show that $(\mathbf{I} - \gamma\mathbf{T})^{-1} = L$.

   Solution:

(a) i. Under the assumption that $\|\mathbf{x}\|_\infty \le 1$, we have

$$\begin{aligned}\|\mathbf{Mx}\|_\infty &= \max_i \left|\sum_{j=1}^n m_{i,j}x_j\right| \le \max_i \sum_{j=1}^n |m_{i,j}||x_j| \\ &\le \max_i \sum_{j=1}^n |m_{i,j}|\|x\|_\infty \le \max_i \sum_{j=1}^n |m_{i,j}|.\end{aligned}$$

Let $i^* \in \mathbf{argmax}_i \sum_{j=1}^n |m_{i,j}|$. The equality holds if $x_j = \operatorname{sgn}(m_{i^*,j})$ for $j = 1, \dots, n$. Therefore $\|\mathbf{M}\|_\infty = \max_{\|\mathbf{x}\|_\infty \le 1} \|\mathbf{Mx}\|_\infty = \max_i \sum_{j=1}^n |m_{i,j}|$.

ii. Using the result in the previous part, we have

$$\begin{aligned}\|\mathbf{AB}\|_\infty &= \max_i \sum_{j=1}^p \sum_{k=1}^n |a_{i,k}b_{k,j}| \le \max_i \sum_{k=1}^n \left(|a_{i,k}| \sum_{j=1}^p |b_{k,j}|\right) \\ &\le \max_i \sum_{k=1}^n |a_{i,k}|\|\mathbf{B}\|_\infty = \|\mathbf{A}\|_\infty \|\mathbf{B}\|_\infty.\end{aligned}$$

(b) For any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that,

$$\forall\, m > n > N, \quad \left\|\sum_{i=n+1}^m (\gamma\mathbf{T})^i\right\|_\infty \le \sum_{i=n+1}^m \left\|(\gamma\mathbf{T})^i\right\|_\infty \le \sum_{i=n}^m \gamma^i \|\mathbf{T}\|_\infty^i = \sum_{i=n+1}^m \gamma^i < \epsilon,$$

where we use the fact that $\sum_{i=0}^k \gamma^i$ is a Cauchy sequence. We see that $\left\|\sum_{i=0}^k (\gamma\mathbf{T})^i\right\|_\infty$ is bounded, because

$$\left\|\sum_{i=0}^k (\gamma\mathbf{T})^i\right\|_\infty \le \left\|\sum_{i=0}^N (\gamma\mathbf{T})^i\right\|_\infty + \left\|\sum_{i=N+1}^\infty (\gamma\mathbf{T})^i\right\|_\infty \le \left\|\sum_{i=0}^N (\gamma\mathbf{T})^i\right\|_\infty + \epsilon,$$

By the Bolzano-Weierstrass theorem, there exists a subsequence of $\left\{\sum_{i=0}^k (\gamma\mathbf{T})^i\right\}_{k=0}^\infty$ that converges to a matrix L. Then, we can find some $K > N$ such that

$$\left\|\sum_{i=0}^K (\gamma\mathbf{T})^i - L\right\| < \epsilon,$$

and thus, for $k > K$, we have

$$\left\|\sum_{i=0}^k (\gamma\mathbf{T})^i - L\right\|_\infty \le \left\|\sum_{i=K+1}^k (\gamma\mathbf{T})^i\right\|_\infty + \left\|\sum_{i=0}^K (\gamma\mathbf{T})^i - L\right\|_\infty < \epsilon + \epsilon = 2\epsilon.$$

Letting $\epsilon \to 0$, we obtain $\lim_{k\to\infty} \sum_{i=0}^k (\gamma\mathbf{T})^i = L$.

(c) $(\mathbf{I} - \gamma\mathbf{T}) \sum_{i=0}^\infty (\gamma\mathbf{T})^i = \sum_{i=0}^\infty (\gamma\mathbf{T})^i - \sum_{i=1}^\infty (\gamma\mathbf{T})^i = (\gamma\mathbf{T})^0 = \mathbf{I}$. So $(\mathbf{I} - \gamma\mathbf{T})^{-1} = L$. ■

Exercise 4: Grid World with a Given Policy

Consider the grid world shown in Figure 7. The finite state space is $\mathcal{S} = \{s_i : i = 1, 2, \ldots, 11\}$ and the finite action space is $\mathcal{A} = \{\text{up, down, left, right}\}$.

State transition probabilities: After the agent picks and performs a certain action, there are four possibilities for the next state: the destination state, the current state, the states to the right and left of the current state. If the states are reachable, the corresponding probabilities are 0.7, 0.1, 0.05, and 0.15, respectively; otherwise, the agent stays where it is. The game will terminate if the agent arrives at s_{10} (loss) or s_{11} (win).

Reward: After the agent picks and performs a certain action at its current state, it receives rewards of 100, -100, and 0, if it arrives at states s_{11}, s_{10}, and all the other states, respectively.

Policy: In Figure 7, the arrows show the policy $\pi : \mathcal{S} \to \mathcal{A}$ for the agent. The random variable S_t is the state at time t under the policy π.

1. Find the matrix $\mathbf{M} \in \mathbb{R}^{11\times 11}$ with $m_{i,j} = \mathbf{P}(S_{t+1} = s_j | S_t = s_i)$, i.e., the conditional probability of the agent moving from s_i to s_j.

   Solution:

$$\mathbf{M} = \begin{pmatrix} 0.25 & 0.7 & 0 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.05 & 0.8 & 0.15 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.05 & 0.25 & 0 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 & 0.7 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3 & 0 & 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.15 & 0 & 0.1 & 0.7 & 0 & 0.05 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.25 & 0.7 & 0 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0.25 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0.7 & 0 & 0 & 0.25 & 0.05 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

   ■

2. Suppose that the initial state distribution is uniform distribution, that is $\mathbf{P}(S_0 = s_i) = 1/11$, $i = 1, \ldots, 11$.

   (a) Find the distributions $\mathbf{P}(S_1)$ and $\mathbf{P}(S_2)$ by following the policy π.

   (b) Show that the agent would finally arrive at either s_{10} or s_{11}, i.e.,

$$\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 9.$$

   (c) Find $\lim_{t\to\infty} \mathbf{P}(S_t = s_{10})$ and $\lim_{t\to\infty} \mathbf{P}(S_t = s_{11})$.

   Solution:

   (a) $\mathbf{P}(S_1) = \mathbf{P}(S_0)\mathbf{M} = \left(\frac{3}{110}, \frac{31}{220}, \frac{2}{55}, \frac{1}{22}, \frac{1}{11}, \frac{3}{22}, \frac{1}{11}, \frac{3}{20}, \frac{3}{110}, \frac{1}{10}, \frac{17}{110}\right)$,
   $\mathbf{P}(S_2) = \mathbf{P}(S_1)\mathbf{M} = \left(\frac{61}{4400}, \frac{147}{1100}, \frac{133}{4400}, \frac{39}{1100}, \frac{29}{550}, \frac{71}{1100}, \frac{553}{4400}, \frac{29}{176}, \frac{3}{220}, \frac{233}{2200}, \frac{571}{2200}\right)$.

(b) The transition probability matrix $\mathbf{M}$ can be written as

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{0,0} & \mathbf{M}_{0,1} \\ \mathbf{O} & \mathbf{I}_2 \end{pmatrix}.$$

Then, the t-step transition probability matrix becomes

$$\mathbf{M}^t = \begin{pmatrix} \mathbf{M}_{0,0}^t & \sum_{k=0}^{t-1} \mathbf{M}_{0,0}^k \mathbf{M}_{0,1} \\ \mathbf{O} & \mathbf{I}_2 \end{pmatrix}.$$

We note that the sum of each row of $\mathbf{M}_{0,0}^4$ is less than 1 (by calculation, or by the observation that either s_{10} or s_{11} can be reached from any state within 4 steps), which implies that $\left\|\mathbf{M}_{0,0}^4\right\|_\infty < 1$, and hence $\lim_{t\to\infty} \mathbf{M}_{0,0}^t = \mathbf{O}$. Moreover, we have $\sum_{k=0}^{\infty} \mathbf{M}_{0,0}^t = (\mathbf{I}_9 - \mathbf{M}_{0,0})^{-1}$. Therefore,

$$\lim_{t\to\infty} \mathbf{M}^t = \begin{pmatrix} \mathbf{O} & (\mathbf{I}_9 - \mathbf{M}_{0,0})^{-1} \mathbf{M}_{0,1} \\ \mathbf{O} & \mathbf{I}_2 \end{pmatrix}.$$

And the limit of the distribution $\mathbf{P}(S_t)$ is

$$\begin{aligned} \lim_{t\to\infty} \mathbf{P}(S_t) &= \mathbf{P}(S_0) \begin{pmatrix} \mathbf{O} & (\mathbf{I}_9 - \mathbf{M}_{0,0})^{-1} \mathbf{M}_{0,1} \\ \mathbf{O} & \mathbf{I}_2 \end{pmatrix} \\ &= \left(0, \ldots, 0, \lim_{t\to\infty} \mathbf{P}(S_t = s_{10}), \lim_{t\to\infty} \mathbf{P}(S_t = s_{11})\right). \end{aligned}$$

(c) According to the previous part, we have

$$\begin{aligned} \left(\lim_{t\to\infty} \mathbf{P}(S_t = s_{10}), \lim_{t\to\infty} \mathbf{P}(S_t = s_{11})\right) &= \mathbf{P}(S_0) \begin{pmatrix} (\mathbf{I}_9 - \mathbf{M}_{0,0})^{-1} \mathbf{M}_{0,1} \\ \mathbf{I}_2 \end{pmatrix} \\ &\approx (0.126446, 0.873554) \end{aligned}$$

■

3. Find the value function corresponding to the policy π, where the discount factor $\gamma = 0.9$.

Solution:
The value function $V^\pi(s)$ is given by $V = (\mathbf{I} - \gamma\mathbf{M})^{-1}R$, where the i-th element of the vector V is $V^\pi(s_i)$, and the i-th element of the vector R is

$$\mathbb{E}\left[r(s_i, \pi(s_i))\right] = \begin{cases} 100 \times m_{i,11} - 100 \times m_{i,10}, & \text{if } s_i \notin \{s_{10}, s_{11}\}, \\ 0, & \text{if } s_i \in \{s_{10}, s_{11}\}. \end{cases}$$

Substituting the data into the above equations, we obtain

$$\begin{aligned} & V = (\mathbf{I} - 0.9\mathbf{M})^{-1}(0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; 0 \;\; -5 \;\; 70 \;\; -5 \;\; -100 \;\; 100)^\top \\ \Longrightarrow \quad & V^\pi(s_1) = 34.216, \;\; V^\pi(s_2) = 38.509, \;\; V^\pi(s_3) = 68.465, \;\; V^\pi(s_4) = 50.159, \\ & V^\pi(s_5) = 81.472, \;\; V^\pi(s_6) = 58.121, \;\; V^\pi(s_7) = 70.290, \;\; V^\pi(s_8) = 94.404, \\ & V^\pi(s_9) = 40.795, \;\; V^\pi(s_{10}) = 0, \;\; V^\pi(s_{11}) = 0. \end{aligned}$$

■

4. Show that the result in (2b) holds for any initial probabilities we choose for $\mathbf{P}(S_0 = s_i)$, $i = 1, \ldots, 11$.

   Solution:
   We have already shown (2b) without any assumption on the initial probabilities. ■

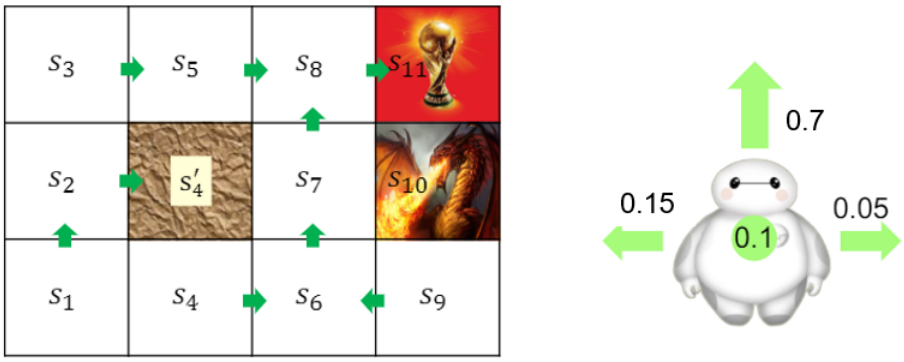

Figure 7: Illustration of a grid world with a policy.

Exercise 5: Optimal Policy

Consider the grid world problem described in Exercise 4. Let π^* be the optimal policy, V^* the corresponding value function, Q^* the corresponding Q function, and $\gamma = 0.9$.

1. Please derive the Bellman optimality equation in terms of the value function V^* and the Q function Q^*, respectively.

   Solution:
   For a given policy $\pi(s)$, we have

$$\begin{aligned} V^\pi(s) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid S_t = s\right] \\ &= \mathbb{E}\left[R_t \mid S_t = s\right] + \gamma \sum_{s'} \mathbf{P}\left(S_{t+1} = s' \mid S_t = s\right) \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_{t+1} = s'\right] \\ &= \mathbb{E}\left[r(s, \pi(s))\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, \pi(s)\right) V^\pi(s'). \end{aligned}$$

   The optimal value function V^* is the the solution to the following Bellman optimality equation

$$\begin{aligned} V^*(s) = \max_\pi V^\pi(s) &= \max_\pi \left\{ \mathbb{E}\left[r(s, \pi(s))\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, \pi(s)\right) V^\pi(s') \right\} \\ &= \max_a \left\{ \mathbb{E}\left[r(s, a)\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, a\right) V^*(s') \right\}. \end{aligned}$$

   The Q function for the given policy $\pi(s)$ is defined as

$$\begin{aligned} Q^\pi(s, a) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k} \mid S_t = s, S_{t+1} = \delta(s, a)\right] \\ &= \mathbb{E}\left[R_t \mid S_t = s\right] + \gamma \sum_{s'} \mathbf{P}\left(S_{t+1} = s' \mid S_{t+1} = \delta(s, a)\right) \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_{t+1} = s'\right] \\ &= \mathbb{E}\left[r(s, a)\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, a\right) V^\pi(s') \\ &= \mathbb{E}\left[r(s, a)\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, a\right) Q^\pi(s', \pi(s')). \end{aligned}$$

   The optimal Q function Q^* is the the solution to the following Bellman optimality equation

$$\begin{aligned} Q^*(s, a) = \max_\pi Q^\pi(s, a) &= \mathbb{E}\left[r(s, a)\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, a\right) \max_\pi V^\pi(s') \\ &= \mathbb{E}\left[r(s, a)\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, a\right) V^*(s') \\ &= \mathbb{E}\left[r(s, a)\right] + \gamma \sum_{s'} \mathbf{P}\left(s' \mid s, a\right) \max_{a'} Q^*(s', a'). \end{aligned}$$

   ■

2. Please choose one of the algorithms we introduced in class to find π^* and V^* respectively and write their pseudocode (hand in your code if you have one).

   Solution:
   For consistency with the lecture notes, we denote the transition matrix as $\mathbf{T}$ instead of $\mathbf{M}$. The pseudocode for the value iteration algorithm is as follows

Algorithm 1 Value Iteration

Input: The transition probabilities $\mathbf{P}(s' \mid s, a)$, the reward function $r(s, a)$, the discount rate γ, the initial value V^0 and the tolerance $\epsilon > 0$.
Output: The optimal value function V^* and the optimal policy π^*.
$k \leftarrow 0$
loop
  $V^{k+1} \leftarrow V^k$, $k \leftarrow k + 1$
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $Q(s, a) \leftarrow \mathbb{E}[r(s, a)] + \gamma \sum_{s'} \mathbf{P}(s' \mid s, a) V^k(s')$
    $V^k(s) \leftarrow \max_a Q(s, a)$, $\pi(s) \leftarrow \arg\max_a Q(s, a)$
  if $\max_s |V^k(s) - V^{k-1}(s)| < \epsilon$ **then**
    $V^* \leftarrow V^k$, $\pi^* \leftarrow \pi$
    break

   The pseudocode for the policy iteration algorithm is as follows

Algorithm 2 Policy Iteration

Input: The transition probabilities $\mathbf{P}(s' \mid s, a)$, the reward function $r(s, a)$, the discount rate γ, the initial policy π^0 and the tolerance ϵ.
Output: The optimal value function V^* and the optimal policy π^*.
$\pi \leftarrow \pi^0$
loop
  $\pi' \leftarrow \pi$
  for $s \in \mathcal{S}$ **do**
    $R^\pi(s) = \mathbb{E}[r(s, \pi(s))]$
    for $s' \in \mathcal{S}$ **do**
      $\mathbf{T}^\pi(s, s') \leftarrow \mathbb{E}[\mathbf{P}(s' \mid s, \pi(s))]$
  $V^\pi \leftarrow (\mathbf{I} - \gamma \mathbf{T}^\pi)^{-1} R^\pi$
  for $s \in \mathcal{S}$ **do**
    for $a \in \mathcal{A}$ **do**
      $Q^\pi(s, a) \leftarrow \mathbb{E}[r(s, a)] + \gamma \sum_{s'} \mathbf{P}(s' \mid s, a) V^\pi(s')$
    $\pi(s) \leftarrow \arg\max_a Q^\pi(s, a)$
  if $\pi = \pi'$ **then**
    $V^* \leftarrow V$, $\pi^* \leftarrow \pi$
    break

   The pseudocode for the Q-learning algorithm is as follows

Algorithm 3 Q-Learning

Input: The discount rate γ, the initial Q function Q^0 and the number of episodes N.
Output: The optimal value function V^* and the optimal policy π^*.
$\hat{Q} \leftarrow Q^0$, $n \leftarrow 0$
for $i = 1$ to N **do**
  Start a new episode at the state s
  while $s \neq$ terminal state **do**
    $a \leftarrow \epsilon$-greedy$(\hat{Q}(s, \cdot))$
    Take the action a, observe the reward r and the state s'
    $\alpha = \frac{1}{n(s,a)+1}$
    $\hat{Q}(s,a) \leftarrow \hat{Q}(s,a) + \alpha \left[r + \gamma \max_{a'} \hat{Q}(s', a') - \hat{Q}(s,a) \right]$
    $n(s,a) \leftarrow n(s,a) + 1$, $s \leftarrow s'$
$V^* \leftarrow \max_a \hat{Q}(\cdot, a)$, $\pi^* \leftarrow \arg\max_a \hat{Q}(\cdot, a)$

A Python implementation of the Q-learning algorithm is given in "HW7.ipynb". ■

3. Please design a reward scheme such that following the resulting optimal policy will never lose. Specifically, you need to derive the resulting optimal policy (the proof is not required) and show

$$\lim_{t \to \infty} \mathbf{P}(S_t = s_i) = 0,\ i = 1, \ldots, 10,$$

whenever $\mathbf{P}(S_0 = s_{10}) = \mathbf{P}(S_0 = s_{11}) = 0$.

Solution:
In order not to lose, we set $r(s,a) = -\infty$ if and only if $\delta(s,a) = s_{10}$. Futhermore, to ensure that the agent will finally reach s_{11} under the optimal policy, we let $r(s,a) > 0$ if $\delta(s,a) = s_{11}$. Otherwise, $r(s,a) = 0$. To sum up, the reward function can be formulated as

$$r(s,a) = \begin{cases} -\infty, & \text{if } \delta(s,a) = s_{10}, \\ 1, & \text{if } \delta(s,a) = s_{11}, \\ 0, & \text{otherwise.} \end{cases}$$

As a result, we have

- A policy under which s_{10} is reachable from the state s has a value $V^\pi(s) = -\infty$.
- A policy under which both s_{10}, s_{11} are unreachable from s has a value $V^\pi(s) = 0$.
- A policy under which s_{10} is unreachable and s_{11} is reachable from s has a value $V^\pi(s) > 0$.

The optimal policy is the one that maximizes the value function and hence satisfies the last statement above. Note that such policy is not unique, so we use the Q-learning algorithm implemented in "HW7.ipynb" to find the exact optimal policy, which is given by

$$\pi^*(s_1) = \text{up}, \quad \pi^*(s_2) = \text{up}, \quad \pi^*(s_3) = \text{right}, \quad \pi^*(s_4) = \text{left}, \quad \pi^*(s_5) = \text{right},$$
$$\pi^*(s_6) = \text{left}, \quad \pi^*(s_7) = \text{left}, \quad \pi^*(s_8) = \text{right}, \quad \pi^*(s_9) = \text{down}.$$

The transition matrix is then

$$\mathbf{M} = \begin{pmatrix} 0.25 & 0.7 & 0 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.3 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.05 & 0.25 & 0 & 0.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0.3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.3 & 0 & 0 & 0.7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.7 & 0 & 0.25 & 0.05 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.15 & 0.8 & 0.05 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.05 & 0.25 & 0 & 0 & 0.7 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0 & 0 & 0.95 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Following the the same framework of proof as in Exercise 4.2(b), we write the transition matrix as

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_1 & \mathbf{0} & \mathbf{M}_2 \\ \mathbf{0} & 1 & 0 \\ \mathbf{0} & 0 & 1 \end{pmatrix}, \quad \mathbf{M}^t = \begin{pmatrix} \mathbf{M}_1^t & \mathbf{0} & \sum_k^{t-1} \mathbf{M}_1^k \mathbf{M}_2 \\ \mathbf{0} & 1 & 0 \\ \mathbf{0} & 0 & 1 \end{pmatrix}.$$

By noting that $\|\mathbf{M}_1^6\|_\infty < 1$, we conclude that

$$\lim_{t\to\infty} \mathbf{M}^t = \begin{pmatrix} \mathbf{O} & \mathbf{0} & (\mathbf{I}-\mathbf{M}_1)^{-1}\mathbf{M}_2 \\ \mathbf{0} & 1 & 0 \\ \mathbf{0} & 0 & 1 \end{pmatrix},$$

which implies that

$$\lim_{t\to\infty} \mathbf{P}(S_t = s_i) = \begin{cases} 0, & \text{if } i = 1, \ldots, 9, \\ \mathbf{P}(S_0 = s_{10}) = 0, & \text{if } i = 10, \\ \mathbf{P}(S_0) \begin{pmatrix} (\mathbf{I}-\mathbf{M}_1)^{-1}\mathbf{M}_2 \\ 0 \\ 1 \end{pmatrix}, & \text{if } i = 11, \end{cases}$$

whenever $\mathbf{P}(S_0 = s_{10}) = 0$. ■

Exercise 6: Value Iteration and Policy Iteration

Consider a Markov Decision Process with bounded rewards and finite state-action pairs. The transition probability is $\mathbf{P}(s'|s,a)$, the discounted factor is $\gamma \in (0,1)$, and the reward function is $r(s,a)$. Let $\pi : \mathcal{S} \to \mathcal{A}$ be a deterministic policy and $Q^\pi(s,a)$ be the accumulated reward by performing the action a first and then following the policy π.

1. Let V^k denote the value function after the k-th iteration of the value iteration algorithm. Please show that value iteration achieves linear convergence rate, that is

$$\left\|V^* - V^{k+1}\right\|_\infty \leq \gamma \left\|V^* - V^k\right\|_\infty.$$

Solution:
For simplicity, we consider the value iteration where the value function is updated by

$$V^{k+1}(s) = \max_a \left\{ \mathbb{E}\left[r(s,a)\right] + \gamma \sum_{s'} \mathbf{P}(s'|s,a) V^k(s') \right\},$$

instead of the one in Algorithm 1. By the Bellman Equation, we have

$$\begin{aligned}
&\left|V^*(s) - V^{k+1}(s)\right| \\
=\ &\left| \max_a \left\{ \mathbb{E}\left[r(s,a)\right] + \gamma \sum_{s'} \mathbf{P}(s'|s,a) V^*(s') \right\} - \max_a \left\{ \mathbb{E}\left[r(s,a)\right] + \gamma \sum_{s'} \mathbf{P}(s'|s,a) V^k(s') \right\} \right| \\
\leq\ &\max_a \left| \gamma \sum_{s'} \mathbf{P}(s'|s,a) V^*(s') - \gamma \sum_{s'} \mathbf{P}(s'|s,a) V^k(s') \right| \\
=\ &\gamma \max_a \sum_{s'} \mathbf{P}(s'|s,a) \left|V^*(s') - V^k(s')\right| \\
\leq\ &\gamma \max_a \sum_{s'} \mathbf{P}(s'|s,a) \left\|V^* - V^k\right\|_\infty = \gamma \|V^* - V^k\|_\infty,
\end{aligned}$$

for any $s \in \mathcal{S}$. Therefore, $\left\|V^* - V^{k+1}\right\|_\infty \leq \gamma \left\|V^* - V^k\right\|_\infty$. ■

2. (a) Find the Bellman Equation for Q^π.
   (b) Consider a new policy π' given by

$$\pi'(s) = \mathbf{argmax}_{a \in \mathcal{A}}\ Q^\pi(s,a).$$

Note that if $\mathbf{argmax}_{a\in\mathcal{A}}\, Q^\pi(s,a)$ is not unique, we can choose one action arbitrarily. Show that $V^{\pi'}(s) \geq V^\pi(s)$ for all $s \in \mathcal{S}$.

Solution:

(a) As is shown in Exercise 5.1, the Bellman Equation is

$$Q^\pi(s,a) = \mathbb{E}\left[r(s,a)\right] + \gamma \sum_{s'} \mathbf{P}(s'|s,a) Q^\pi(s', \pi(s')).$$

(b) By the definition of $\pi'(s)$, we have

$$\begin{aligned} V^{\pi'}(s) - V^{\pi}(s) &= Q^{\pi'}(s, \pi'(s)) - Q^{\pi}(s, \pi(s)) \geq Q^{\pi'}(s, \pi'(s)) - Q^{\pi}(s, \pi'(s)) \\ &= \gamma \sum_{s'} \mathbf{P}(s'|s, \pi'(s)) \left[Q^{\pi'}(s', \pi(s')) - Q^{\pi}(s', \pi(s')) \right] \\ &= \gamma \sum_{s'} \mathbf{P}(s'|s, \pi'(s)) \left[V^{\pi'}(s') - V^{\pi}(s') \right]. \end{aligned}$$

Or equivalently, in the language of matrix,

$$\left(\mathbf{I} - \gamma \mathbf{T}^{\pi'} \right) \left(V^{\pi'} - V^{\pi} \right) \geq \mathbf{0}.$$

Since $\left(\mathbf{I} - \gamma \mathbf{T}^{\pi'} \right)^{-1} = \sum_{k=0}^{\infty} \gamma \mathbf{T}^{\pi'}$ has non-negative entries, multiplying both sides of the inequality by it does not change the sign. Therefore, we obtain $V^{\pi'} - V^{\pi} \geq \mathbf{0}$, i.e. $V^{\pi'}(s) \geq V^{\pi}(s)$ for all $s \in \mathcal{S}$. ■

Exercise 7: Q-learning algorithm (Optional)

1. Consider the Q-learning algorithm for any deterministic MDP with finite state-action pairs and non-negative rewards. Assume that we initialize all $\hat{Q}$ values to zero. Let $\hat{Q}_n(s,a)$ denote the learned $\hat{Q}(s,a)$ value after the n-th iteration of the training procedure in Q learning algorithm.

   (a) Please show that $\hat{Q}$ values never decrease during training, that is

   $$\hat{Q}_{n+1}(s,a) = r(s,a) + \gamma \max_{a'} \hat{Q}_n(s',a') \geq \hat{Q}_n(s,a),\ \forall s,a,n,$$

   where s' is the state the agent attains after taking action a at state s.

   (b) Please show that throughout the training process, every $\hat{Q}$ value will remain in the interval between zero and the optimal Q function Q^*, that is

   $$0 \leq \hat{Q}_n(s,a) \leq Q^*(s,a),\ \forall s,a,n.$$

   Solution:

   (a) Since $\hat{Q}_0(s,a) = 0$ and $r(s,a) \geq 0$, we have

   $$\hat{Q}_1(s,a) - \hat{Q}_0(s,a) = r(s,a) > 0,\ \forall s,a.$$

   Assume that $\hat{Q}_n(s,a) \geq \hat{Q}_{n-1}(s,a)$, for all s,a and some $n \geq 1$. Then, we have

   $$\begin{aligned}
   \hat{Q}_{n+1}(s,a) - \hat{Q}_n(s,a) &= \gamma \left(\max_{a'} \hat{Q}_n(s',a') - \max_{a'} \hat{Q}_{n-1}(s',a') \right) \\
   &= \gamma \left(\max_{a'} \hat{Q}_n(s',a') - \hat{Q}_{n-1}(s',a^*) \right) \\
   &\geq \gamma \left(\hat{Q}_n(s',a^*) - \hat{Q}_{n-1}(s',a^*) \right) \geq 0,\ \forall s,a,
   \end{aligned}$$

   where we let $a^* \in \mathbf{argmax}_{a'} \hat{Q}_{n-1}(s',a')$. By induction on n, $\hat{Q}_{n+1}(s,a) - \hat{Q}_n(s,a) \geq 0$ holds for all $n \geq 0$, i.e. $\hat{Q}_{n+1}(s,a) \geq \hat{Q}_n(s,a)$, $\forall s,a,n$.

   (b) The first inequality holds immediately from the fact that $\hat{Q}_n(s,a) \geq \hat{Q}_0(s,a) = 0$. To see the second inequality, we again perform induction on n. The base case $n = 0$ is trivial, as $r(s,a) \geq 0$. Assume that $Q^*(s,a) \geq \hat{Q}_{n-1}(s,a)$ for all s,a and some $n \geq 1$. Then, we have

   $$\begin{aligned}
   \hat{Q}^*(s,a) - \hat{Q}_n(s,a) &= \gamma \left(\max_{a'} \hat{Q}^*(s',a') - \max_{a'} \hat{Q}_{n-1}(s',a') \right) \\
   &= \gamma \left(\max_{a'} \hat{Q}^*(s',a') - \hat{Q}_{n-1}(s',a^*) \right) \\
   &\geq \gamma \left(\hat{Q}^*(s',a^*) - \hat{Q}_{n-1}(s',a^*) \right) \geq 0,\ \forall s,a,
   \end{aligned}$$

   where we let $a^* \in \mathbf{argmax}_{a'} \hat{Q}_{n-1}(s',a')$. In conclusion, $\hat{Q}^*(s,a) \geq \hat{Q}_n(s,a)$, $\forall s,a,n$. ■

2. Consider the Q-learning algorithm for a stochastic MDP with finite state-action pairs. The transition probability is $\mathbf{P}(s'|s,a)$ and the reward function is deterministic, denoted by $r(s,a)$. Assume that we initialize all $\hat{Q}$ values to zero. Let $\hat{Q}_n(s,a)$ denote the learned $\hat{Q}(s,a)$ value after the n-th iteration of the training procedure in Q-learning algorithm. Please show that

$$\mathbb{E}_{s'\sim\mathbf{P}(\cdot|s,a)}\left[r(s,a)+\gamma\max_{a'}\hat{Q}_n(s',a')\right]\geq r(s,a)+\gamma\max_{a'}\mathbb{E}_{s'\sim\mathbf{P}(\cdot|s,a)}\left[\hat{Q}_n(s',a')\right],\forall s,a,n.$$

Solution:
Since the reward function is deterministic, we have

$$\mathbb{E}_{s'\sim\mathbf{P}(\cdot|s,a)}\left[r(s,a)+\gamma\max_{a'}\hat{Q}_n(s',a')\right]=r(s,a)+\gamma\sum_{s'}\mathbf{P}(s'|s,a)\max_{a'}\hat{Q}_n(s',a').$$

Thus, we only need to show that

$$\sum_{s'}\mathbf{P}(s'|s,a)\max_{a'}\hat{Q}_n(s',a')\geq\max_{a'}\sum_{s'}\mathbf{P}(s'|s,a)\hat{Q}_n(s',a'),$$

Let $a^*\in\mathbf{argmax}_{a'}\sum_{s'}\mathbf{P}(s'|s,a)\hat{Q}_n(s',a')$. Then, the above inequality becomes

$$\sum_{s'}\mathbf{P}(s'|s,a)\max_{a'}\hat{Q}_n(s',a')\geq\sum_{s'}\mathbf{P}(s'|s,a)\hat{Q}_n(s',a^*),$$

which is clearly true. ■