EECS 127 / EE 227AT 9/27/18
L. El Ghaoui, A. Bayen

Homework Assignment #3

Due date: 10/11/18, before class. Please LaTeX or handwrite your homework solution and submit an electronic version.

Exercise 1 (Kantorovich Problem)

Question: How to move mass μ to ν with minimal cost?

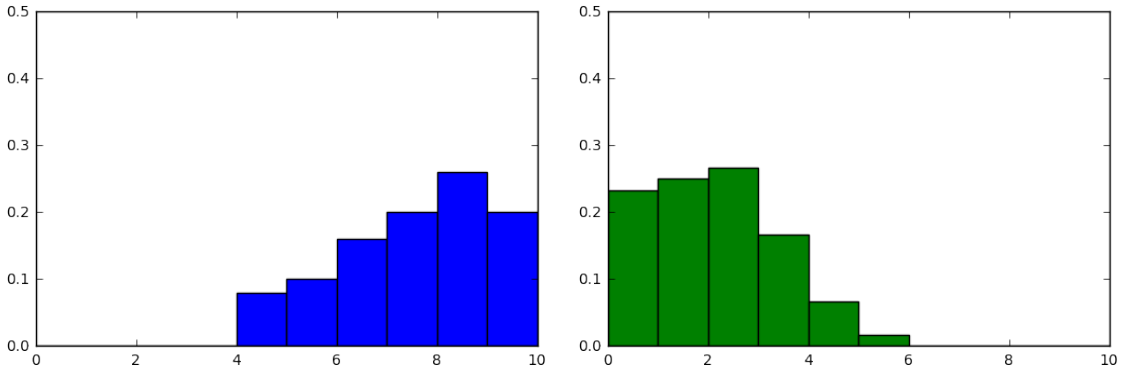

Figure 1: Visualization of μ histogram on left and ν histogram on right.

More rigorously, let $n \in \mathbb{N}$. We define two discrete probability distributions $\mu = (\mu_1, \cdots, \mu_n)$ and $\nu = (\nu_1, \cdots, \nu_n)$ with $\sum_i \mu_i = \sum_i \nu_i = 1$. We define $C = (c_{ij})_{\substack{1\cdots n \\ 1\cdots n}} \in \mathbb{R}_+^{n^2}$ be the cost matrix for transporting one unit of mass from location $i \in [1, \cdots, n]$ to location $j \in [1, \cdots, n]$. The flow matrix $P = (p_{ij})_{\substack{1\cdots n \\ 1\cdots n}}$ denotes the quantity of mass to be moved from location i to location j. P satisfies the limit conditions:

$$P \mathbb{1}_n = \mu$$
$$P^T \mathbb{1}_n = \nu$$

1. What is the total cost of transporting the mass μ into ν by following the transportation plan P?
2. Write the optimization problem of finding the transportation plan P with minimal cost. What type of optimization problem is it? (LP, QP, $\cdots$?).

The optimal value of the optimization problem you found is a well-defined distance named *Kantorovich* or Wasserstein distance noted $\mathcal{W}(\mu, \nu)$. That distance can be readily applied in the context of document similarity. In particular, while computating the Kantorovich between two documents, we obtain a flow matrix. The point of this question is to show how this matrix can be interpreted. For that, we provide you with a certain prior distance on words (word embeddings) that can be used to compute the transportation cost in the notebook `text-kantorovich.ipynb`.

3. Calculate the Wasserstein distance in the notebook using CVX and use visualize the resulting flow matrix P using the provided code. Interpret the results.

4. Compute the cosine distance between the two given documents. Comment on which distance is more insightful on this precise example and why.

   Note: Cosine distance between two documents (or sets of words) is computed as follows. Collect ordered sets of all word counts from both documents. Then compute the cosine of the angle between the two vectors. For example from "black blue black" and "blue blue red", one would obtain the sets $\{black : 2, blue : 1, red : 0\}$ *and* $\{black : 0, blue : 2, red : 1\}$. *We can then compute the cosine distance between the vectors* $[2, 1, 0]$ *and* $[0, 2, 1]$.

We are now interested in computing different barycenters between two discrete vectors μ and ν shown in Figure 1. Let us note that the convex combination $t\mu + (1-t)\nu$ has the variational formulation $\operatorname{argmin}_x t(\mu - x)^2 + (1-t)(\nu - x)^2$. Inspired from this fact, we define the variational formulation for convex combinations on the Wasserstein space $\operatorname{argmin}_x t\mathcal{W}(x, \mu) + (1-t)\mathcal{W}(x, \nu)$.

5. Write this optimization problem as a LP.

6. Solve the optimization problem using CVX in the notebook `barycenter.ipynb`. Visualize the Waserstein barycenter for $t \in [0, 0.25, 0.5, 0.75, 1.0]$.

7. Use your results to comment on the differences between Wasserstein convex combination and euclidean convex combination.

Exercise 2 (Fast CV for Least-Squares) In this exercise, we consider a regularized least-squares problem:

$$w_\lambda := \arg\min_w \ \|y - Xw\|_2^2 + \lambda\|w\|_2^2, \tag{1}$$

with $X \in \mathbb{R}^{n\times p}$ the data matrix (with one data point per row), $y \in \mathbb{R}^n$ is the response vector, and $\lambda > 0$ a "ridge" regularization parameter. Solving the above problem leads to a prediction model: for a new data point $x \in \mathbb{R}^p$, $\hat{y}(x) = w_\lambda^\top x$.

We would like to choose this regularization parameter based on the notion of "leave-one-out" (LOO) cross-validation, whereby for a given candidate value of $\lambda > 0$, we estimate the resulting prediction error, averaged across all the models given by the above, when leaving out one data point. Precisely, we set, for $i = 1, \ldots, n$

$$w_\lambda^{(i)} := \arg\min_w \ \|y_{\setminus i} - X_{\setminus i} w\|_2^2 + \lambda\|w\|_2^2,$$

with $X_{\setminus i}$ (resp. $y_{\setminus i}$) is equal to X (resp. y), with the i-th row (resp. element) removed, and evaluate the prediction error on the point we just left out, with $\hat{y}(x_i) = x_i^\top w_\lambda^{(i)}$.

Obviously, we can compute the LOO error by simply solving n problems of the form (1), with the appropriate data. In this exercise we investigate a faster method, which is based on solving the above full problem *once*, then performing cheap updates to get the LOO error.

1. Show that the solution to the full problem is of the form

$$w_\lambda = (X^\top X + \lambda I)^{-1} X^\top y$$

2. Prove a (simple) version of the Sherman-Morrison-Woodbury identity. Given $M = A + uv^\top$ (with A symmetric/invertible, $A + uv^\top$ also invertible) show that

$$M^{-1} = A^{-1} - \frac{1}{1 + v^\top A^{-1} u}(A^{-1}u)(A^{-1}v)^\top \tag{2}$$

   Namely, it is sufficient to just show that $MM^{-1} = I$ (since showing $M^{-1}M = I$ is similar).

3. Argue both $X_{\setminus i}^\top X_{\setminus i}$ and $X_{\setminus i}^\top y_{\setminus i}$ can be written as rank-one modifications of $X^\top X$ and $X^\top y$ and show that,

$$w_\lambda^{(i)} = w - \frac{\Sigma^{-1} x_i e_i}{1 - h_i}$$

   where we define $\Sigma = X^\top X + \lambda I$, $h_i = x_i^\top \Sigma^{-1} x_i$, and $e_i = y_i - x_i^\top w_\lambda$ for convenience. *Hint: use the Sherman-Morrison-Woodbury identity.*

4. Compute (and simplify) the LOO prediction error $\frac{1}{n}\sum_{i=1}^n (y_i - (w_\lambda^{(i)})^\top x_i)^2$ into an expression consisting of e_i, h_i.

5. What is the complexity of the method for computing the LOO prediction error investigated in this problem relative the naive method where we compute all $w_\lambda^{(i)}$ without reusing computations? Highlight the leading dependencies in terms of n, p for both methods in big-O notation.

Exercise 3 (Least norm estimation on traffic flow networks) In this problem, we want to estimate the traffic given the road network as well as the historical average of flows on each road segment. We call q_i the flow of vehicles on each road segment $i \in I$. At each intersection, the sum of all incoming flows must be equal to the sum of all outgoing flows. We construct the matrix $A \in \mathbb{R}^{J \times I}$ such that the element on the jth line and ith column is

- 0 if link i does not arrive or leave intersection j;
- 1 if link i arrives at intersection j;
- -1 if link i leaves intersection j.

1. Write down the linear equation that corresponds to the conservation of vehicles at each intersection $j \in J$.
2. The goal is to estimate the traffic flow on each of the road segment. The flow estimates should satisfy the conservation of vehicles exactly at each intersection. Among the solutions that satisfy this constraint, we are searching for the estimate that is the closest to the historical average, $\overline{q}$, in the l_2-norm sense. The vector $\overline{q}$ has size I and the i-th element represent the average for the road segment i. Pose the optimization problem.
3. Find a closed form solution to this problem. Detail your answer (do not only give a formula but explain where it comes from).

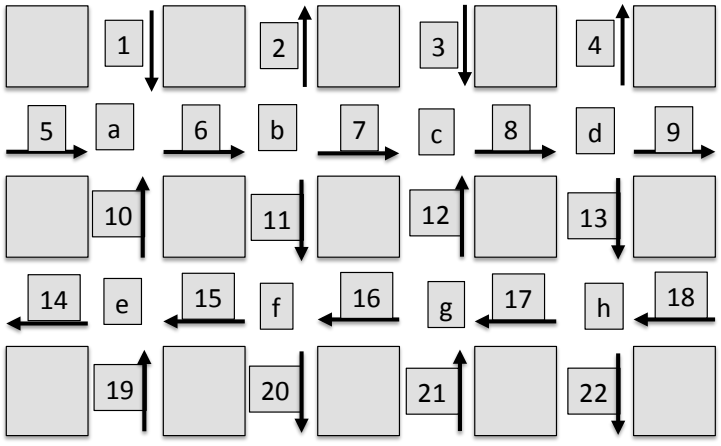

Figure 2: Example of traffic estimation problem. The intersections are labeled a to h. The road segments are labeled 1 to 22. The arrows indicate the direction of traffic.

4. Formulate the problem for the small example of Figure 2 and solve it using the historical average given in Table 1. What is the flow that you estimate on road segments 1, 3, 6, 15 and 22?

segment	average	measured
1	2047.6	2028
2	2046.0	2008
3	2002.6	2035
4	2036.9	
5	2013.5	2019
6	2021.1	
7	2027.4	
8	2047.1	
9	2020.9	2044
10	2049.2	
11	2015.1	
12	2035.1	
13	2033.3	
14	2027.0	2043
15	2034.9	
16	2033.3	
17	2008.9	
18	2006.4	
19	2050.0	2030
20	2008.6	2025
21	2001.6	
22	2028.1	2045

Table 1: Table of flows: historical averages $\overline{q}$ (center column), and some measured flows (right column).

5. Now, assume that besides the historical averages, you are also given some flow measurements on some of the road segments of the network. You assume that these flow measurements are correct and want your estimate of the flow to match these measurements perfectly (besides matching the conservation of vehicles of course). The right column of Table 1 lists the road segments for which we have such flow measurements. Do you estimate a different flow on some of the links? Give the difference in flow you estimate for road segments 1,3, 6, 15 and 22. Also check that you estimate gives you the measured flow on the road segments for which you have measured the flow. *Hint:* Your solution will comment on the feasibility of solving such a problem.

Exercise 4 (A Portfolio Design Problem) The returns on $n = 4$ assets are described by a Gaussian (normal) random vector $r \in \mathbb{R}^n$, having the following expected value $\hat{r}$ and covariance matrix Σ:

$$\hat{r} = \begin{bmatrix} 0.12 \\ 0.10 \\ 0.07 \\ 0.03 \end{bmatrix}, \ \Sigma = \begin{bmatrix} 0.0064 & 0.0008 & -0.0011 & 0 \\ 0.0008 & 0.0025 & 0 & 0 \\ -0.0011 & 0 & 0.0004 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The last (fourth) asset corresponds to a risk-free investment. An investor wants to design a portfolio mix with weights $x \in \mathbb{R}^n$ (each weight x_i is non-negative, and the sum of the weights is one) so as to obtain the best possible expected return $\hat{r}^T x$, while guaranteeing that:

(i) No single asset weights more than 40%
(ii) The risk-free assets should not weight more than 20%
(iii) No asset should weight less than 5%
(iv) The probability of experiencing a return lower than $q = -3\%$ should be no larger than $\epsilon = 10^{-4}$

1. What is the maximal achievable expected return, under the above constraints?

   Hint: Constraint (iv) is known as a "chance constraint."

   $$a^T x \sim N(\hat{a}, \Sigma) \implies a^T x - b \sim N(\hat{a}x - b, x^T x \Sigma x).$$

   We then have:

   $$\Pr(a^T x \leq b) \geq \eta \iff b - \hat{a}^T x \geq \Phi^{-1}(\eta)||\Sigma^{1/2} x||_2$$

2. Solve the problem for a large number of values of ϵ between 10^{-4} and 10^{-1}, and plot the optimal values of $\hat{r}^T x$ versus ϵ. Also make an area plot of the optimal portfolios x versus ϵ.

3. *Monte Carlo simulation.* Let x be the optimal portfolio found in part 1, with $\epsilon = 10^4$. This portfolio maximizes the expected return, subject to the probability of a loss being no more than 3 %. Generate 1000 samples of r, and plot a histogram of the returns. Find the empirical mean of the return samples, and calculate the percentage of samples for which a loss occurs.

Exercise 5 (A slalom problem) A two-dimensional skier must slalom down a slope, by going through n parallel gates of known position (x_i, y_i), and of width c_i, $i = 1, \ldots, n$. The initial position (x_0, y_0) is given, as well as the final one, (x_{n+1}, y_{n+1}). Here, the x-axis represents the direction down the slope, from left to right, see Figure 3.

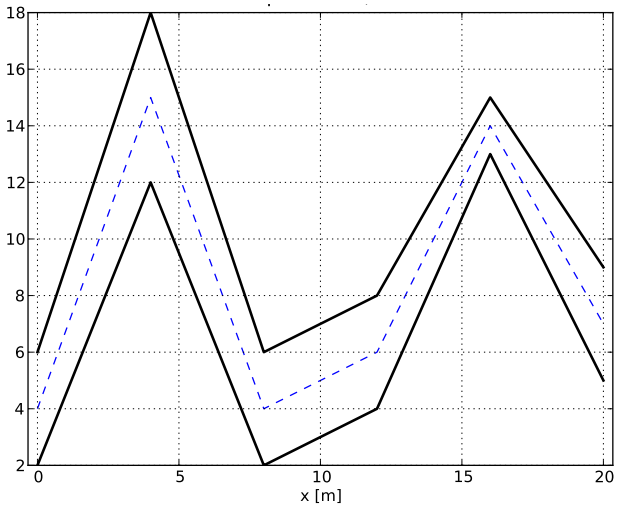

Figure 3: Slalom problem with $n = 5$ obstacles. "Uphill" (resp. "downhill") is on the left (resp. right) side. The middle path is dashed, initial and final positions are not shown.

i	x_i	y_i	c_i
0	0	4	N/A
1	4	5	3
2	8	4	2
3	12	6	2
4	16	5	1
5	20	7	2
6	24	4	N/A

Table 2: Problem data for Exercise 5.

1. Find the path that minimizes the total length of the path. Your answer should come in the form of an optimization problem.

2. Solve the problem numerically, with the data given in Table 2.