IFT 6390 HW1 Solutions

Zhaoliang Yang (with minor edits by instructor and TAs)

September 2019

1. **Solution to Q1**

   (a) The conditional probability of discrete random variable X is that the probability that X occurs given the knowledge that variable Y has occurred. This conditional probability is written as $P(X|Y)$. If events A and B are not independent, the probability that both events happen is calculated by $P(XandY) = P(X|Y)P(Y)$. Usually $P(XandY)$ is written as $P(X \cap Y)$. Therefore the equation for conditional probability is

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)}$$

   (b) The 3 tosses of biased coin are mutually independent. Suppose event A is 2 heads and 1 tail during 3 tosses, event B is head up and apparently $P(B) = 2/3$. $P(A \cap B)$ is

$$P(A \cap B) = \frac{2}{3} \times C_2^1 \times \frac{2}{3} \times \frac{1}{3} = \frac{8}{27}$$

   According to conditional probability equation mentioned above,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{8/27}{2/3} = \frac{4}{9}$$

   (c) Solutions to (c) are as follows:

   (i) $P(X, Y) = \mathbb{P}(Y|X) \cdot \mathbb{P}(X)$

   (ii) $P(X, Y) = \mathbb{P}(X|Y) \cdot \mathbb{P}(Y)$

   (d) As demonstrated in (c), the definition of conditional probability enable us to have the following two equations:

$$P(X, Y) = \mathbb{P}(Y|X) \cdot \mathbb{P}(X) \tag{1}$$

$$P(X, Y) = \mathbb{P}(X|Y) \cdot \mathbb{P}(Y) \tag{2}$$

   Since left sides of equations (1) and (2) are equal, the right sides should be equal too. Therefore we can get a new equations

$$\mathbb{P}(Y|X) \cdot \mathbb{P}(X) = \mathbb{P}(X|Y) \cdot \mathbb{P}(Y) \tag{3}$$

   If both sides of equation (3) are divided by $\mathbb{P}(Y)$ (suppose $\mathbb{P}(Y) \neq 0$), we can get

$$\mathbb{P}(X|Y) = \frac{\mathbb{P}(Y|X) \cdot \mathbb{P}(X)}{\mathbb{P}(Y)} \tag{4}$$

   Equation (4) is Bayes Theorem.

(e) Solutions to (e) are as follows:

(i) As stated in the question, the student is drawn randomly from the surveyed group which contains 55% of UdeM students and 45% of McGill students. Therefore the probability that the student is affliated with McGill is 0.45.

(ii) We can define the following events and their probability.

- **Event A**-Student is from McGill $P(A) = 0.45$
- **Event B**-Student is Bilingual $P(B) = 0.8 \times 0.55 + 0.5 \times 0.45 = 0.665$
- $P(B|A)$ is the probability that the student is bilingual given that the student is from McGill, $P(B|A) = 0.5$

Based on Bayes Theorem, we have

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} = 0.3383$$

2. **Solution to Q2**

(a) from Table 1, it is obvious that

$$P(word = "goal"|topic = politics) = \frac{7}{1000}$$

(b) The expected times n is

$$\begin{aligned} n &= Number\_of\_Words \times P(word = "goal"|topic = sports) \\ &= 200 \times 0.01 \\ &= 2 \end{aligned}$$

(c) The document is either sports or politics and thus the total probability is

$$\begin{aligned} P(word = "goal") &= P(word = "goal"|topic = sports) \times P(sports) \\ &\quad + P(word = "goal"|topic = politics) \times P(politics) \\ &= 0.01 \times 2/3 + 0.007 \times 1/3 \\ &= 0.009 \end{aligned}$$

(d) We define the following events

- **Event A**-a random word from a document is "kick". Similar to (c), we can get $P(A) = 0.005 \times 2/3 + 0.003 \times 1/3 = 13/3000$
- **Event B**-the topic of the document is sports. As stated in the question, $P(B) = 2/3$
- As stated in the question, $P(A|B) = 1/200$

Based on Bayes Theorem,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{(1/200) \times (2/3)}{13/3000} = \frac{10}{13}$$

(e) We define the following events

- **Event A**- draw a random word from a document, given the word is "kick", the document is sports. According to (d), $P(A) = \frac{10}{13}$. $P(A^{'})$ is the probability that given the random word is "kick" but the document is not sports but politics, it is obvious that $P(A\prime) = 1 - P(A) = \frac{3}{13}$.
- **Event B**- draw a random word from a document and the word is "$goal$". $P(B|topic = sport)$ and $P(B|topic = politics)$ are listed in Table 1

$$\begin{aligned} P(second = \text{"}goal\text{"}|first = \text{"}kick\text{"}) &= P(word = \text{"}goal\text{"}|topic = sports) \times P(A) \\ &+ P(word = \text{"}goal\text{"}|topic = politics) \times P(A^{'}) \\ &= (1/100) \times (10/13) + (7/1000) \times (3/13) \\ &= 121/13000 \end{aligned}$$

(f) I would use the frequency that an event happens to estimate its probability. The details are shown below.

- **topic probability**- count the number of documents in each topic. Suppose the number of documents in sports is N_{sports} and the number of documents in politics is $N_{politics}$. It is obvious that $N_{sports} + N_{politics} = N_{documents}$.

$$P(topic = sports) \approx \frac{N_{sports}}{N_{documents}}$$

$$P(topic = politics) \approx \frac{N_{politics}}{N_{documents}}$$

- **conditional probability**-take sports for example. Count each word in all sports documents. $N_{i,word}$ is used to denote the number of specific word in ith document whose topic is sports. The following matrix is achieved after counting. Any word's conditional probability in sports is estimated as

$$P(word|topic = sports) \approx \frac{\sum_{i=1}^{N_{sports}} N_{i,word}}{\sum N_{ij}}$$

$\sum N_{ij}$ is the total of all elements of below matrix. Same method is used to calculate $P(word|topic = politics)$.

$$\begin{bmatrix} N_{1,\text{"}goal\text{"}} & N_{1,\text{"}kick\text{"}} & N_{1,\text{"}congress\text{"}} & N_{1,\text{"}vote\text{"}} & N_{1,other} \\ N_{2,\text{"}goal\text{"}} & N_{2,\text{"}kick\text{"}} & N_{2,\text{"}congress\text{"}} & N_{2,\text{"}vote\text{"}} & N_{2,other} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ N_{N_{sports},\text{"}goal\text{"}} & N_{N_{sports},\text{"}kick\text{"}} & N_{N_{sports},\text{"}congress\text{"}} & N_{N_{sports},\text{"}vote\text{"}} & N_{N_{sports},other} \end{bmatrix}$$

3. **Solution to Q3**

(a) Since $x_1, x_2, \cdots, x_n$ are independently and identically distributed, their joint density distribution can be written as,

$$f_\theta(x_1, x_2, \cdots, x_n) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \cdots f_\theta(x_n) = \prod_{i=1}^{n} f(x_i) \tag{5}$$

(b) As stated in the question, if x is within $[0, \theta]$, $f_\theta(x_i) = 1/\theta, (i = 1, 2, \cdots, n)$. Given equation (5), we can get,

$$f_\theta(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \frac{1}{\theta^n}$$

Otherwise, $f_\theta(x) = 0$ if x is out of $[0, \theta]$. Therefore,

$$f_\theta(x_1, x_2, \cdots, x_n) = \prod_{i=1}^{n} f(x_i) = \begin{cases} \frac{1}{\theta^n} & 0 \leq \mathbf{x} \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

The derivative of equation (6) is,

$$(\frac{1}{\theta^n})' = -\frac{n}{\theta^{n+1}} < 0, \quad \theta \in (0, \infty) \tag{7}$$

Equation (7) shows that equation (6) is a monotonically decreasing function for $\theta > 0$ and that means the smallest value of θ generates the largest/target Maximum Likelihood Estimation. Therefore the smallest value of θ is the target θ_{MLE} that maximizes the likelihood.
On the other hand, as defined by pdf, θ is not smaller than $x_i, i = [1, 2, \cdots, n]$, which means

$$\theta \geqslant max(x_1, x_2, \cdots, x_n) \tag{8}$$

As stated before, θ_{MLE} is the minimum of θ which is $max(x_1, x_2, \cdots, x_n)$ according to equation (8). Therefore we have proved that

$$\theta_{MLE} = max(x_1, x_2, \cdots, x_n)$$

4. **Solution to Q4**
Since $x_1, x_2, \cdots, x_n$ are iid, equation (5) still applies. Because $y = ln(x)$ is a monotonically increasing function for all $x > 0$ with no stationary points, the θ that makes maximum value of $f_\theta(x_1, x_2, \cdots, x_n)$ is the same θ that makes maximum value of $ln[f_\theta(x_1, x_2, \cdots, x_n)]$.

$$\begin{aligned} ln[f_\theta(x_1, x_2, \cdots, x_n)] &= \sum_{i=1}^{n} ln[f(x_i)] \\ &= nln2 + nln\theta + \sum_{i=1}^{n} lnx_i - \theta \sum_{i=1}^{n} x_i^2 \end{aligned} \tag{9}$$

Please note that in equation (9), all are constant except θ. Suppose

$$a = nln2 + \sum_{i=1}^{n} lnx_i, \quad b = \sum_{i=1}^{n} x_i^2$$

equation (9) can be simplified as

$$ln[f_\theta(x_1, x_2, \cdots, x_n)] = a + nln\theta - b\theta = g(\theta) \tag{10}$$

The derivative of $g(\theta)$ is

$$g'(\theta) = \frac{n}{\theta} - b$$

There's **only one** value that makes $g^{'}(\theta) = 0$ and it's $\theta_{g'=0} = n/b$. In addition, if $\theta < n/b$, $g^{'}(\theta) > 0$ while if $\theta > n/b$, $g^{'}(\theta) < 0$. That being said $g(\theta)$ reaches its maximum value at $\theta = n/b$. As stated before, $f_\theta(x_1, x_2, \cdots, x_n)$ will also reaches at its maximum value which is the maximum likelihood. In conclusion, the maximum likelihood of θ is

$$\theta_{MLE} = \frac{n}{b} = \frac{n}{\sum_{i=1}^{n} x_i^2}$$

5. **Solution to Q5**

   (a) In this case, the error will occur if the number of nearest neighbours in the wrong category is larger than that in the correct category. That being said, $P_n(e)$ equals the probability that $j \leqslant (k-1)/2$. j denotes the number of nearest neighbours in D that are in the correct category, $j \in [0, (k-1)/2]$. Pick j from 0 to $(k-1)/2$, we can get

$$\begin{aligned} P_n(e) &= (\frac{1}{2})^0(\frac{1}{2})^n\binom{n}{0} + (\frac{1}{2})^1(\frac{1}{2})^{n-1}\binom{n}{1} + \cdots + (\frac{1}{2})^n(\frac{1}{2})^0\binom{n}{(k-1)/2} \\ &= \frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} \end{aligned} \tag{11}$$

   (b) if $k = 1$, it is very obvious that only the first polynomial element is left in equation (11). Since the rest of the polynomial elements are positive, all $k > 1$ error rate will be larger than $k = 1$.

   (c) One property for combination function $\binom{n}{m}$ is that its value increases until m reaches the middle of n. In this case, $j \in [0, (k-1)/2]$ and $k \leqslant n$ and that being said, $j = (k-1)/2$

makes the largest value of $\binom{n}{j}$. Thus

$$\begin{aligned}
\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} &\leqslant \frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{(k-1)/2} \\
&= \frac{1}{2^n}\frac{k+1}{2}\frac{n!}{(\frac{k-1}{2})!(n-\frac{k-1}{2})!} \\
&= \frac{1}{2^n}\frac{k+1}{2(\frac{k-1}{2})!}\frac{n!}{(n-\frac{k-1}{2})!} \\
&\textbf{Note: } \frac{k+1}{2(\frac{k-1}{2})!} < 1 \text{ for } k > 7 \\
&\leq \frac{1}{2^n}\frac{n!}{(n-\frac{k-1}{2})!} \\
&\leq \frac{1}{2^n}\frac{n!}{(n-k)!} \\
&= \frac{1}{2^n}n(n-1)(n-2)\cdots(n-k+1) \\
&\leq \frac{n^k}{2^n} \\
&\textbf{Note: } k \leq a\sqrt{n} \\
&\leq \frac{n^{a\sqrt{n}}}{2^n} \\
&= \left(\frac{n^a}{2\sqrt{n}}\right)^{\sqrt{n}}
\end{aligned}$$

When $k \in \{1,3,5\}$, we can use $\frac{k+1}{2(\frac{k-1}{2})!} \leq 2$, and the same reasoning holds to show that $\frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} \leq 2\left(\frac{n^a}{2\sqrt{n}}\right)^{\sqrt{n}}$.

It is known that exponential function grows the fastest when n is large enough. Thus $\lim_{n\to\infty}\left(\frac{n^a}{2\sqrt{n}}\right) = 0$, and $\left(\frac{n^a}{2\sqrt{n}}\right)^{\sqrt{n}} \to 0$ when $n \to \infty$. Hence $P_n(e) \to 0$ when $n \to \infty$.

6. **Solution to Q6** According to the statement in the question, we can define the following events and their corresponding probabilities:

   - **Event A**-The probability that $Y = 0$. Since the coin is balanced, $P(A) = 0.5$.
   - **Event B**-The probability that $X = \mathbf{x}$. $\mathbf{x}$ is 50% from $N_2(\mu_{1,1})$ and 50% $N_2(\mu_{2,2})$, thus $P(B) \sim 0.5f_{\mu_1,\Sigma_1} + 0.5f_{\mu_2,\Sigma_2}$.
   - As stated in the question, $P(B|A) \sim N_2(\mu_1, \Sigma_1)$.

According to Bayes Theorem,

$$
\begin{aligned}
P(Y=0|X=\mathbf{x}) &= \frac{P(X=\mathbf{x}|Y=0)P(Y=0)}{P(X=\mathbf{x})} \\
&= \frac{P(B|A)P(A)}{P(B)} \\
&= \frac{0.5 f_{\mu_1,\Sigma_1}}{0.5 f_{\mu_1,\Sigma_1} + 0.5 f_{\mu_2,\Sigma_2}} \\
&= \frac{f_{\mu_1,\Sigma_1}}{f_{\mu_1,\Sigma_1} + f_{\mu_2,\Sigma_2}} \\
&= \frac{1}{1 + \frac{f_{\mu_2,\Sigma_2}}{f_{\mu_1,\Sigma_1}}}
\end{aligned}
$$

Note: insert f_{μ_1,Σ_1} and f_{μ_2,Σ_2} here given by the question.

$$
\begin{aligned}
&= \frac{1}{1 + ae^b} \\
a &= \sqrt{\frac{det(\Sigma_1)}{det(\Sigma_2)}} \\
b &= \frac{1}{2}((\mathbf{x}-\mu_1)^T \Sigma_1^{-1}(\mathbf{x}-\mu_1) - (\mathbf{x}-\mu_2)^T \Sigma_2^{-1}(\mathbf{x}-\mu_2))
\end{aligned}
$$

1. **Discussion on Q5**
   h and σ are hyperparameters for Hard Parzen and Soft Parzen respectively. According to figure 1, I have the following findings/discussion.

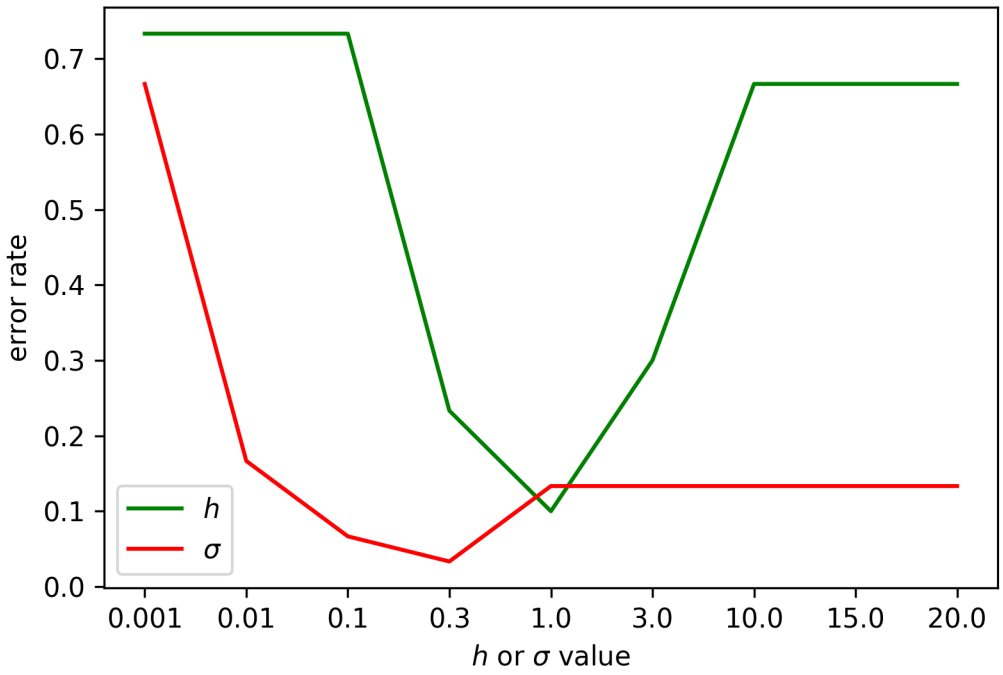

Figure 1: h or σ vs. error rate

(a) With the increase of h and σ, their error rates first decrease and reach their minimal error rates.

(b) The difference is that h stays flat until it reaches 0.1 while σ decreases sharply right at the beginning. h stays flat at the beginning because h is too small to include any neighbours and thus the computation about its neighbours is skipped and the returned result is a random class, which is also the reason why the error rate is so high. The reason why error rate for SoftParzen is high when σ is small is that Gaussian distribution is very sharp/thin if σ is very small and the data points' kernels are too small to be effective.

(c) With further increase of h or σ, their error rates will also increase and finally stay flat when their values are big enough. Error rate will stay the same if h is big enough because all the trained data points has been included as neighbours and the neightbours won't change even if h becomes even larger. In that case, the class is solely determined by number of data points in each class, which is not accurate. When σ is becomes large, Gaussian distribution becomes smooth and relevant kernels are not discriminant. That's why error rate for SoftParzen also becomes large and stay the same when σ is large enough. Overall, SoftParzen has a lower (average) error rate than HardParzen.

(d) A Not-Good h will generate a very high error rate while a Not-Good σ will generate a modest error rate. Overall, SoftParzen has a lower error rate and is more reliable.

(e) h and σ are critical hyperparameters that will impact error rates of Hard Parzen and Soft Parzen models. Plotting empirical lines of error vs. hyperparameter can help us find the hyperparameters that generate the least errors.

2. **Discussion on Q7**

Figure 2 shows how running time will vary with h or σ for each method. y axis is the mean time of 500 running multiplied by 100. It is multiplied by 100 so that the y axis value won't be too small.

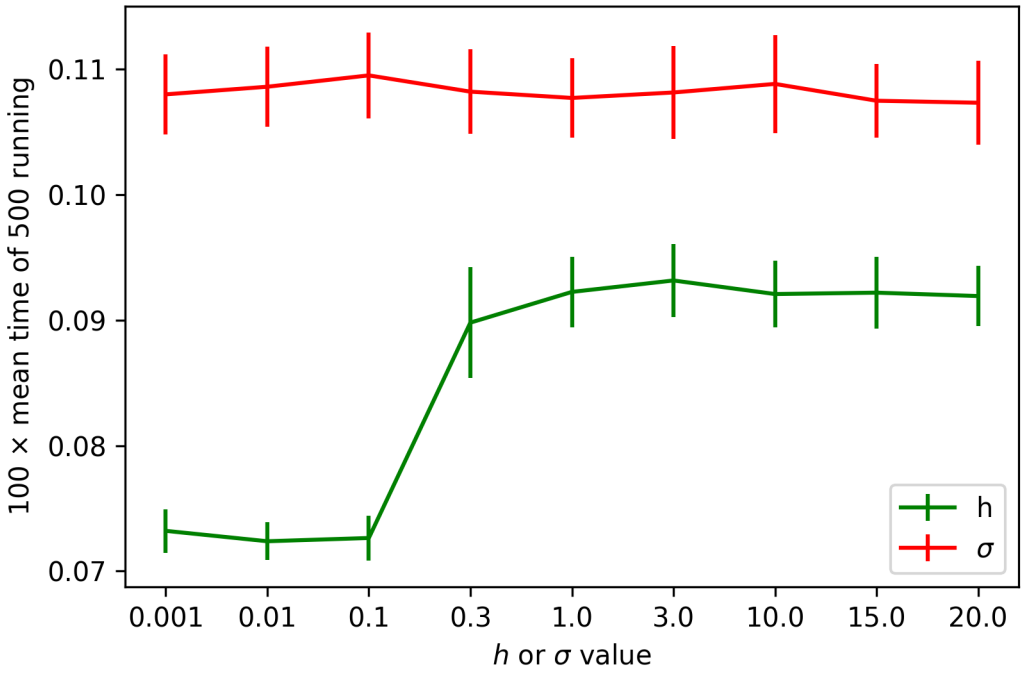

Figure 2: h or σ vs. running time

(a) It is obvious that σ almost stays flat as the SoftParzen's computation doesn't quite vary with the value of σ (All points' kernels will always be calculated). Therefore σ won't affect the running time complexity of SoftParzen.

(b) Running time of HardParzen does vary with h. HardParzen's running time almost stays flat until $h = 0.1$ because h is too small and include zero neighbours and the computation of neighbours' classes and class numbers is skipped. That's why it takes relatively less running time. When h reaches 0.3, the running time increases sharply and doesn't change with further increase of h because their computation processes are the same.

(c) Overall SoftParzen takes more time than HardParzen to compute because SoftParzen has more matrix computation than HardParzen. Additionally, in my functions, SoftParzen has 2 loops and HardPazen only has 1 loop.

3. **Discussion on Q9**
Figure 3 shows how the error rate of random projection methods will change with values of h or σ.I will discuss error rate and running time respectively with previous non-projected methods.

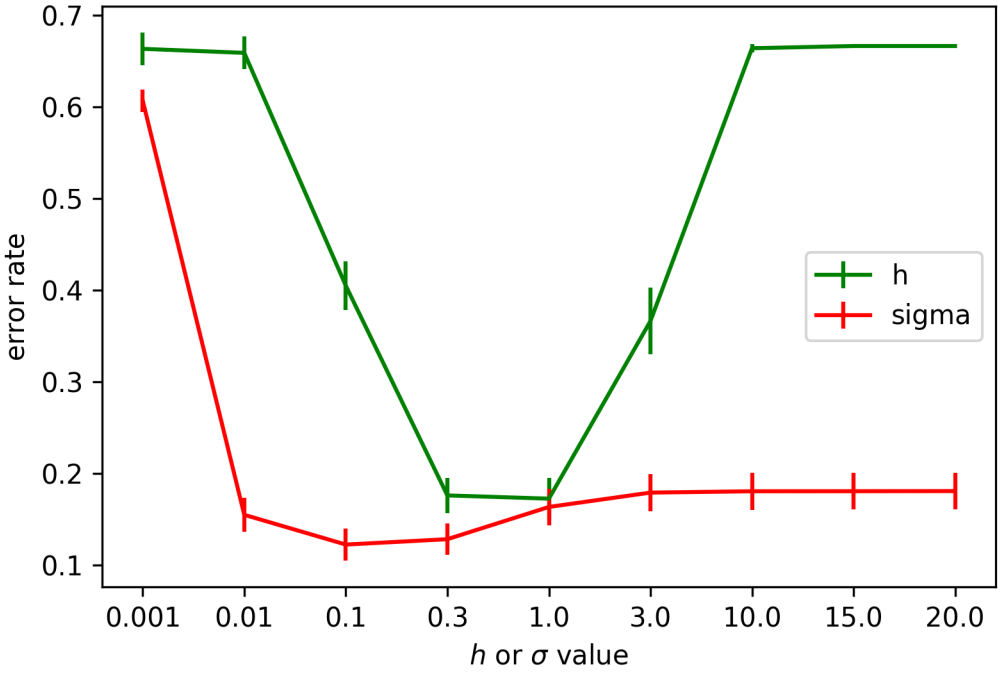

Figure 3: h or σ vs. error rate by projection methods

(a) The basic trend of error rate by projection methods is the same with that of non-projected methods. As analyzed in Q5, the change trend of error rate is determined by the logic of methods/program which has nothing to do with X being projected or not.

(b) The projected methods reaches their minimal error rate at a smaller value of h or σ than non-projected methods. That is because random projection by $A_{4\times2}$ reduced dimensions and make the distance smaller.

(c) The projected methods' minimum error rate is larger than their non-projected counterparts'. My understanding is that some information in X is lost after random projection. The projected methods' minimum error rate is still acceptable though.

(d) On the other hand, it also shows that if accuracy is not that critical, we can use random projection to reduce dimensions, which can be time-saving.

(e) As illustrated by Figure 4 and 5, the basic trend of running time by projection methods is the same with that of non-projected methods. As analyzed in Q7, the change trend of running time is determined by the logic of methods/program which has nothing to do with X being projected or not.

(f) As analyzed in (e), the running time of projected methods is slightly less than that of non-projected methods, as illustrated by fig 5. This is not very obvious though.

(g) in projected methods, SoftParzen still takes more time than HardParzen, which is the same with non-projected methods, because The computation logic of the program doesn't change.

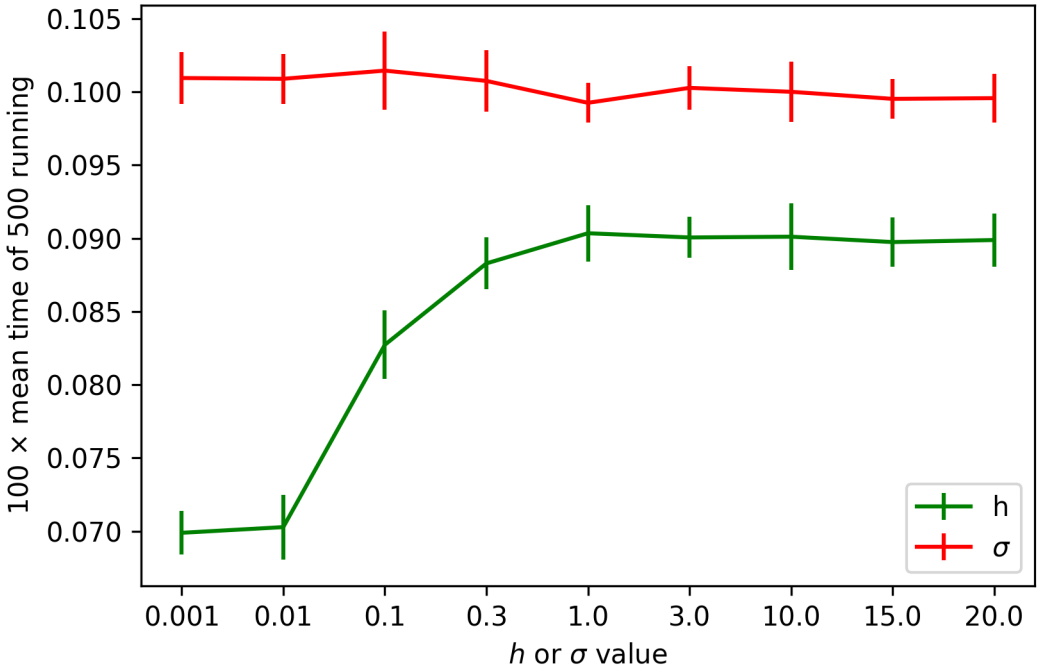

Figure 4: h or σ vs. running time by projection methods

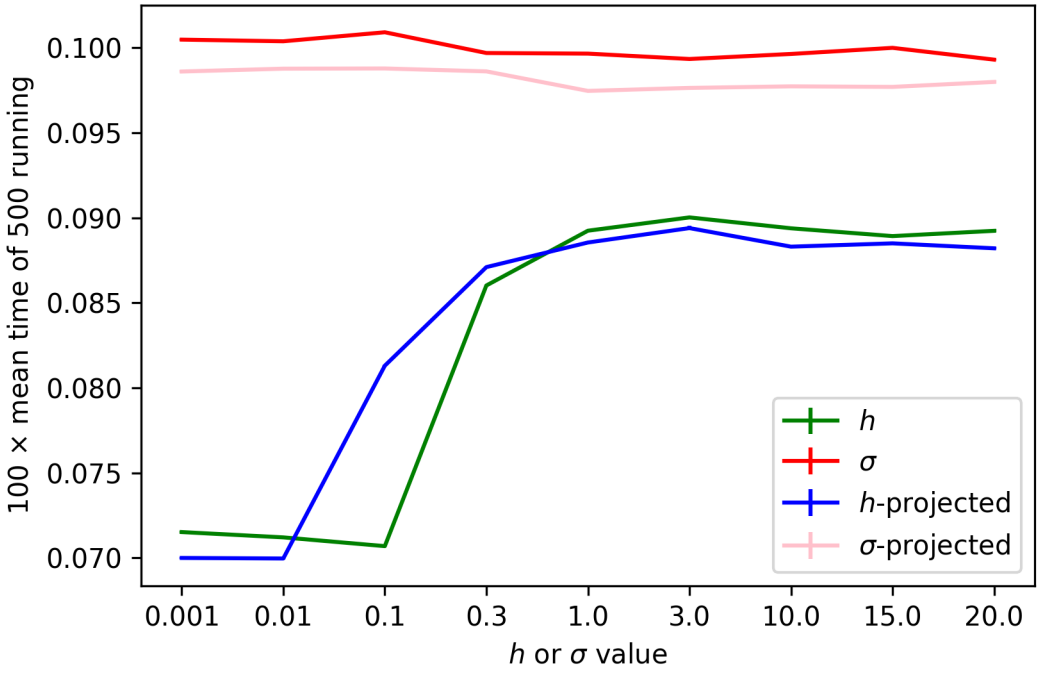

Figure 5: running time comparison-projected vs. non-projected