CS6601 Midterm – Spring 2020

Please read the following instructions thoroughly.

Fill out this PDF form and submit it on **Gradescope** and then on Canvas for backup purposes.

You have unlimited resubmissions until the deadline. You can: **(a)** type directly into the form – we highly recommend using Adobe Reader DC (or Master PDF on Linux). Other programs may not save your answers, so **please keep a backup**; or **(b)** print, hand-write & scan. You can combine the methods as well.

Submit only a single PDF – no phone pictures, please! (You may use an app like CamScanner or Office Lens if you do not have scanner access.) Do not add pages unless absolutely necessary; if you do, please add them at the end of the exam **only**, and clearly label **both** the extra page and the original question page. Submit **ALL** pages of the exam, not only the completed ones.

Do not forget to fill the checklist at the end before turning in the exam. The exam may not be graded if it is left blank.

The exam is open-book, open-note, open video lectures, with no time limit aside from the open period. No internet use is allowed, except for e-text versions of the textbook, this semester's CS6601 course materials, Piazza, and any links provided in the PDF itself. No resources outside this semester's 6601 class should be used. **There is no collaboration on the exams.** Do not discuss the exam on Piazza, Slack, or any other form of communication. If there is a question for the teaching staff, **please make it private on Piazza and tag it as Midterm Exam with the question number in the subject line** (for example, a question on Search would be "Midterm Exam #2 Search").

Please round all your final answers to 6 decimal places, don't round intermediate results. You can use `round(your_number, 6)` function in Python for help.
You will not receive full credit if your answers are not given to the specified precision.

Point breakdown (Each question has sub-parts with varying points):

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Total
Pts	**8**	**12**	**16**	**16**	**16**	**16**	**16**	**100**

1. Game Playing

(8 points)

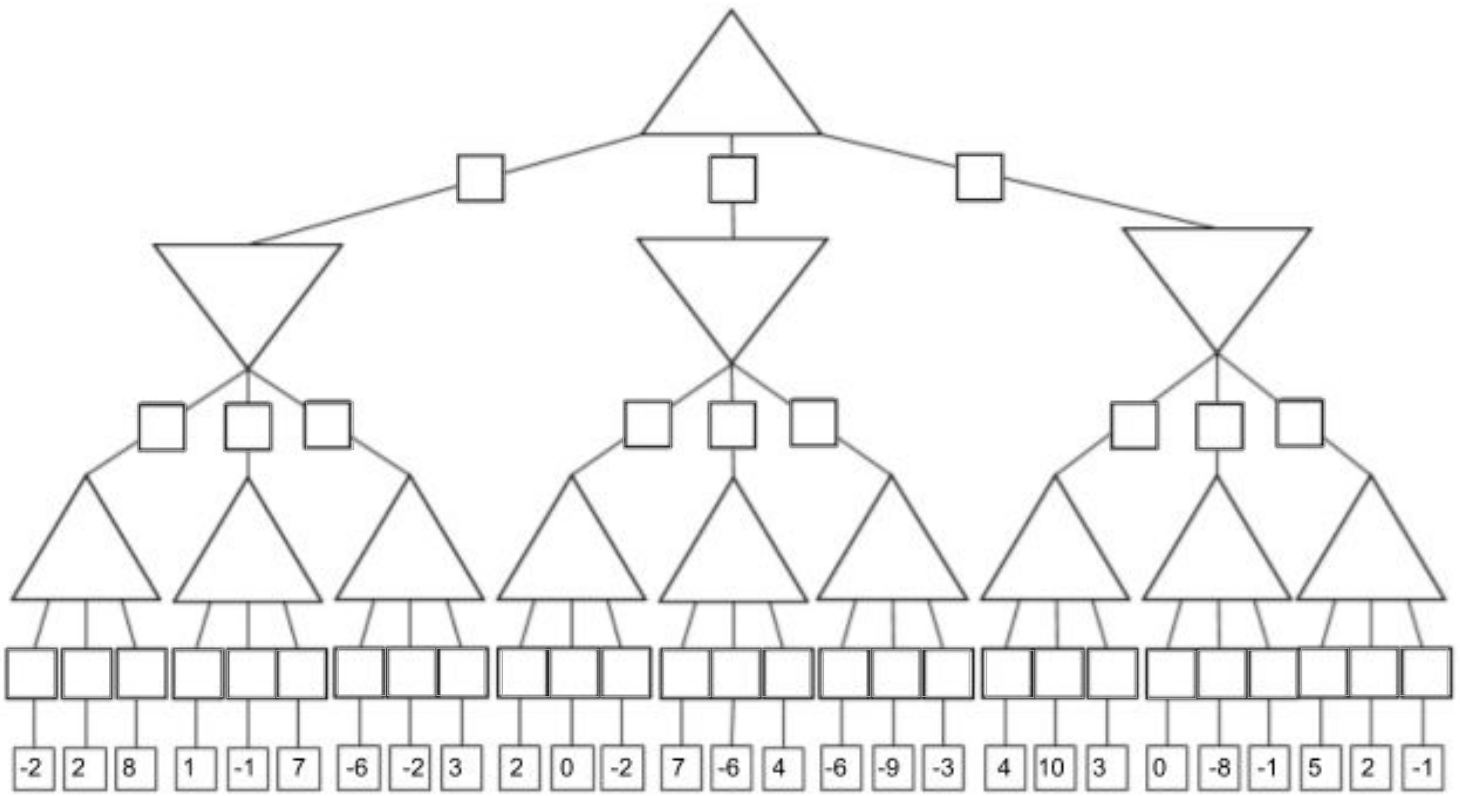
Question A.

(2 points)

Solve the game tree using alpha-beta pruning. The algorithm evaluates the branches from left to right. Indicate which branches are pruned by checking the correct box. Show the value of each node.

When you prune a branch at a top-level, check all the sub boxes/branches down along that path as well. If an upper-level node in a branch gets pruned and you've selected the checkbox to indicate that it is pruned, you DO need to fill in values for the "unvisited" underlying nodes and should also check all the boxes to indicate which branches are pruned.

Make sure to use inequality signs if appropriate.



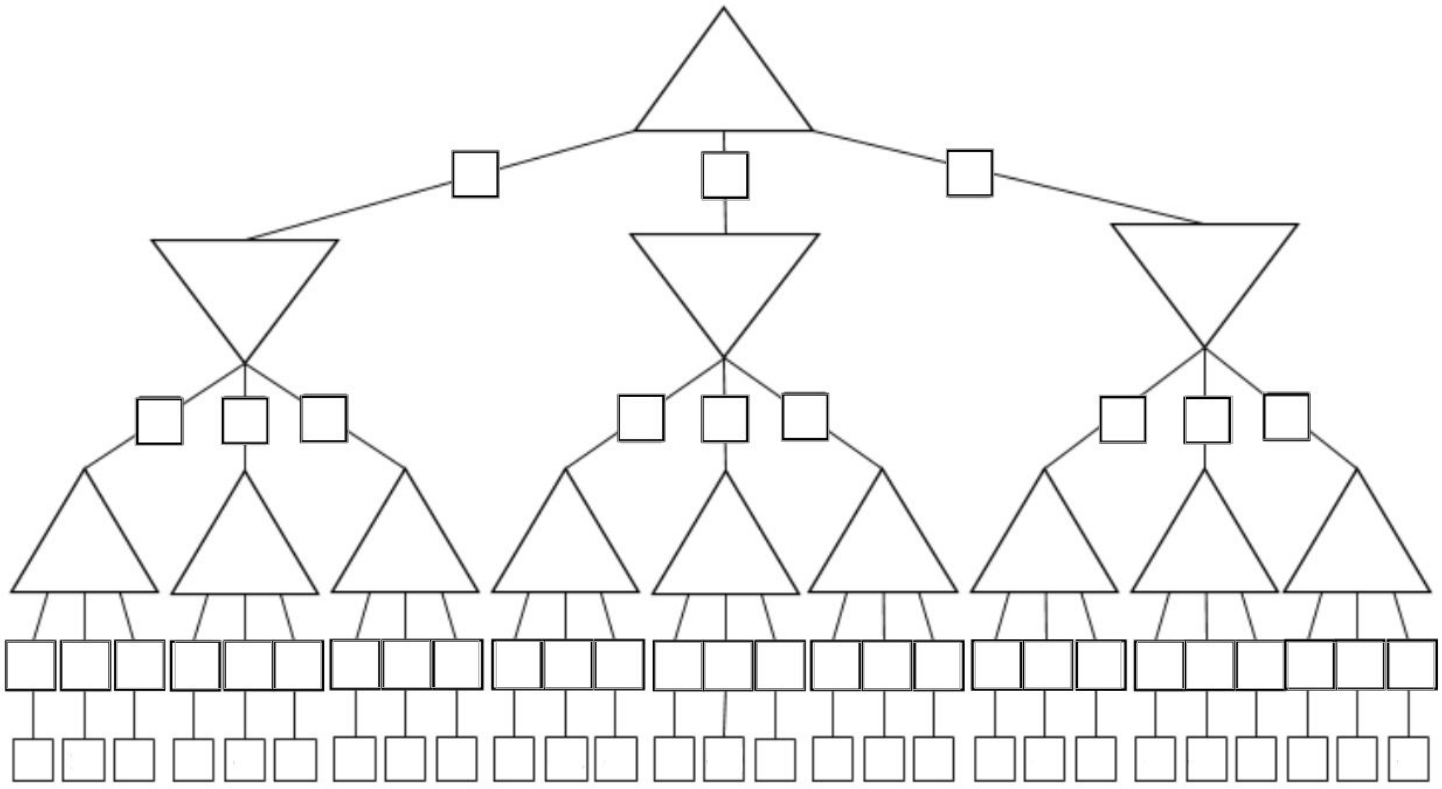
Question B.

(3 points)

Reorder the leaf nodes of the tree from Question A to maximize the number of branches that are pruned. In other words, for the game represented by the tree in Question A, what would the best possible node ordering for Alpha-Beta if the algorithm operates from left to right?

When you prune a branch at a top-level, check all the sub boxes/branches down along that path as well. If an upper-level node in a branch gets pruned and you've selected the checkbox to indicate that it is pruned, you DO need to fill in values for the "unvisited" underlying nodes and should also check all the boxes to indicate which branches are pruned.

Make sure to use inequality signs if appropriate.



B.1. What is the total number of pruned branches?

_____

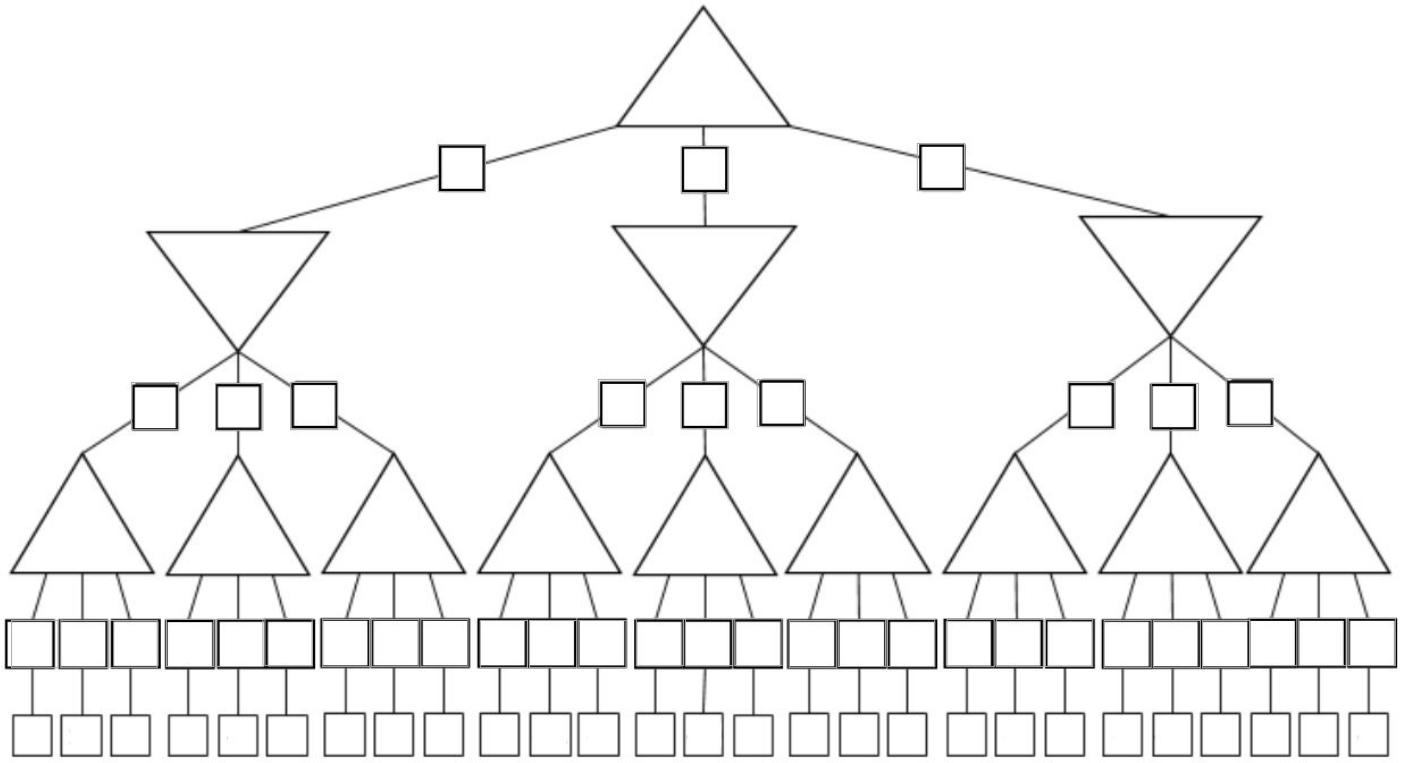
Question C.

(3 points)

Reorder the leaf nodes of the tree from Question A to minimize the number of branches that are pruned. In other words, for the game represented by the tree in Question A, what would the worst possible node ordering for Alpha-Beta if the algorithm operates from left to right?

When you prune a branch at a top-level, check all the sub boxes/branches down along that path as well. If an upper-level node in a branch gets pruned and you've selected the checkbox to indicate that it is pruned, you DO need to fill in values for the "unvisited" underlying nodes and should also check all the boxes to indicate which branches are pruned.

Make sure to use inequality signs if appropriate.



C.1. What is the total number of pruned branches?

_____

2. Search

(12 points)

Traditional search algorithms such as Uniform Cost Search and A* are commonly applied to problem spaces surrounding physical space, but there are many other fields that can take advantage of these algorithms. One such space is the area of optimizing activity in the global currency exchange market. If a particular investor wanted to convert her wealth from the US Dollar to Japanese Yen, it may be more financially beneficial to first exchange their dollars to an intermediary currency, and then exchange that currency to Yen rather than exchange their USD to Yen in one initial exchange. If the state of the current currency exchange market in one moment can be modeled as a graph, then shortest-path algorithms work very successfully to solve this problem. Read the scenario below and work through the following problems to see such an example:

In the distant future, Maks is a middle-class citizen of Isbelland: the nation governed by the powerful political head Charles Isbell. He is a particularly clever investor in the global financial market and has received a tip from a trusted source that the currency of Isbelland, the Isbellion (ISBL), will crash in value over the next couple years. To protect his financial future, he wants to convert his wealth from Isbellions to a much more stable currency that has a stronger future. After sufficient research, he has determined that the currency with the most promising future is the Starner Buck (STNR), the currency of the economic powerhouse Starneria. Isbelland and Starneria have very tense relations, however, and the national governments have imposed a large tax to convert currencies between them. The two nations have a web of political relationships, however, and Maks sees he can utilize intermediary currencies of foreign countries to avoid this tax. He must utilize a combination of currency transfers to transfer his wealth to Starner Bucks, and he realizes this is a problem that he can apply classical search algorithms to.

Firstly, he models the current political and economic landscape as a graph structure, resulting in the following:

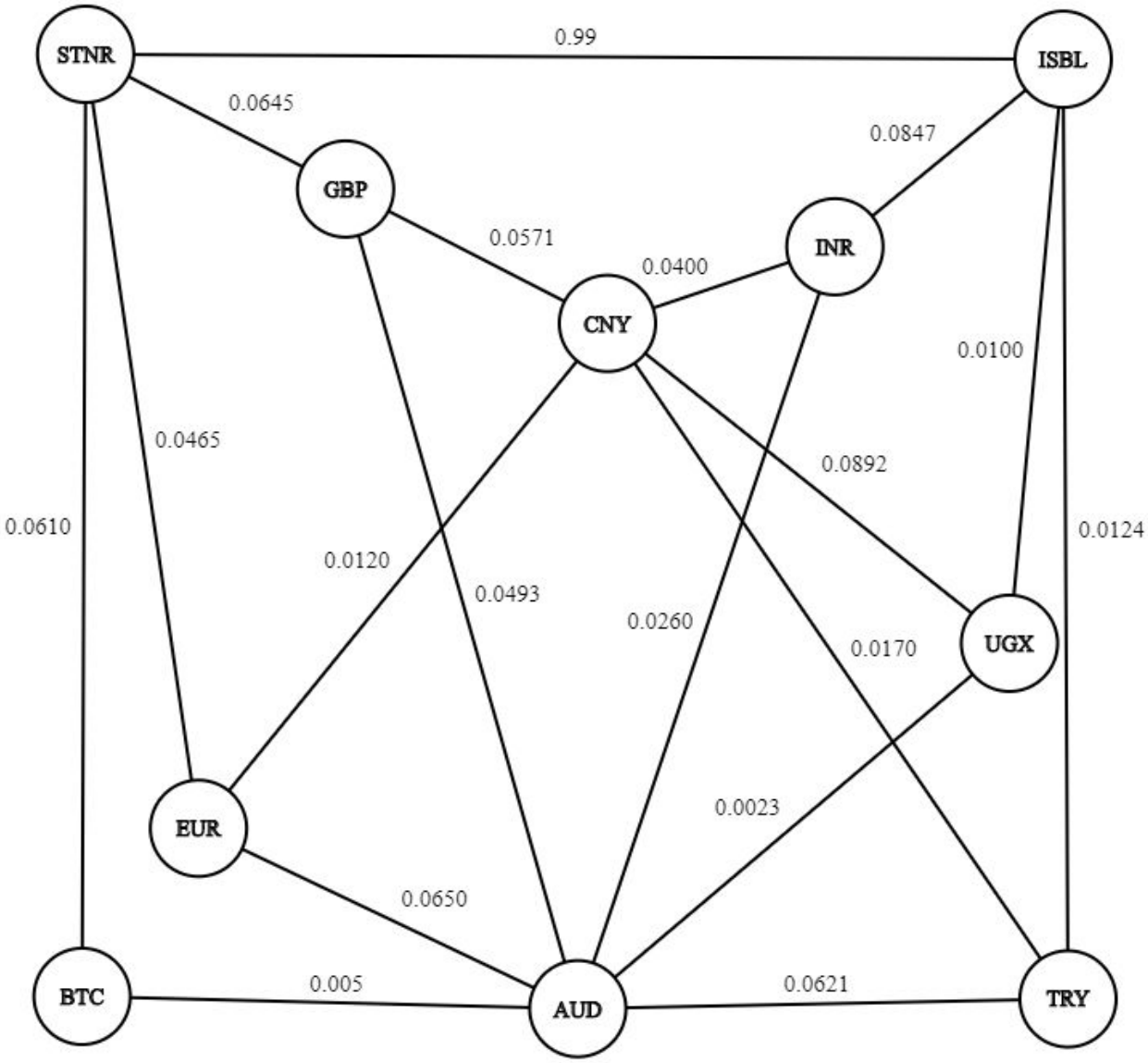


Figure 1.1: Currency Graph

Each node represents a national currency, and each graph edge represents the loss in total value when exchanging wealth between currencies. For example, if Maks was exchanging $100 Australian Dollars (AUD) to Turkish Lira (TRY), the overall value would be reduced by a

factor of .0621 due to transactional costs and he would end up with an amount of Turkish Lira that's actually worth $93.79 AUD.

a. If Maks has a starting value of 10,000 Isbellions and wants to convert all his wealth to Starner Bucks, what's the optimal order of currency exchanges he should perform to lose the least amount of money? Enter your answer as a comma-separated list using the node labels in the graph above. **(3 points)**

_____

b. What is his final value of Starner Bucks when evaluated in Isbellions? **(1 point)**

_____

After performing all his exchanges, Maks is reassured that his wealth is protected despite the falling value of the Isbellion. He becomes slightly curious about his source of information and decides to look further as to the reasons behind the predicted currency devaluation. After researching the topic in-depth, Maks comes to the realization that Isbelland is on the verge of completely failing as a state and descending into anarchy due to gross mismanagement. Wanting to avoid the fallout of this catastrophe, he wants to travel to Starneria and become a citizen as soon as possible. He no longer cares about the financial cost and wants to relocate in the fastest manner possible. Luckily, Isbelland and Starneria are connected by landmass and Maks owns a car. See the information below and answer the following questions about the A* algorithm:

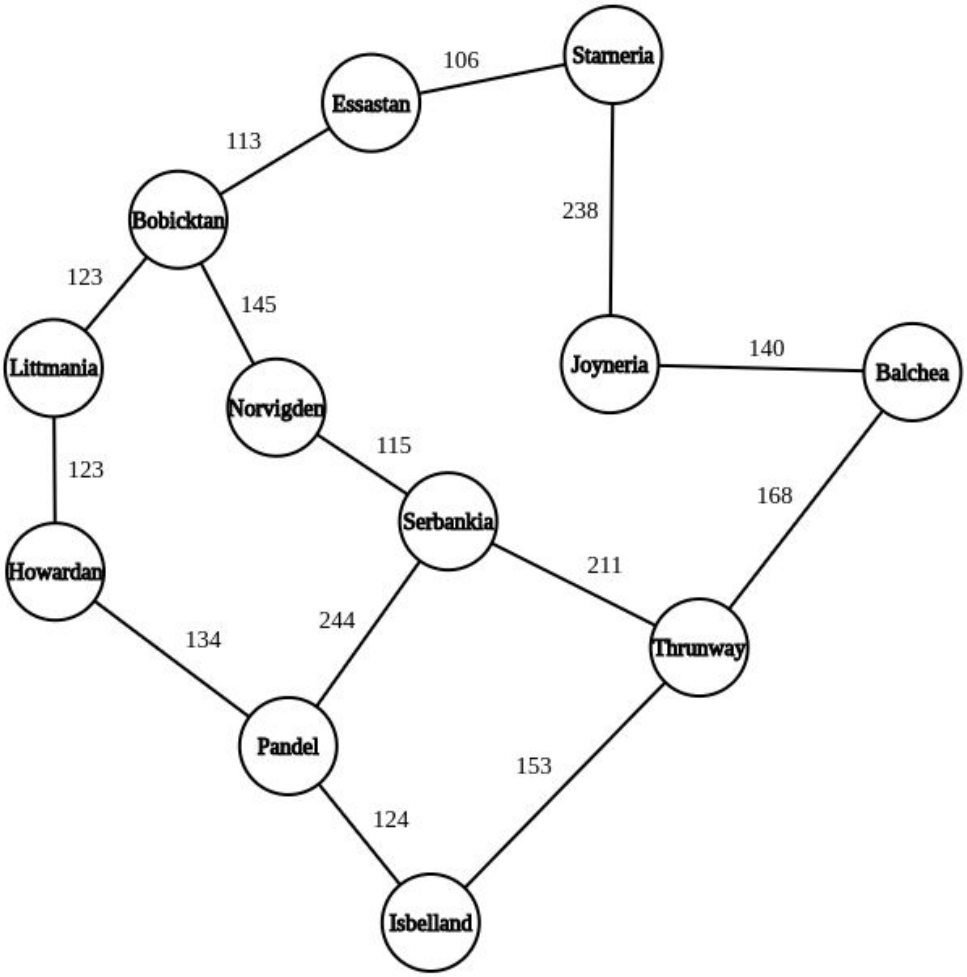


Figure 1.2: Distance Graph

For the sake of simplicity, imagine that Maks never needs to refill his gas tank and travels at a constant speed. Additionally, each geographical location will be represented by a simple (X, Y) coordinate pair. See these coordinates in the table below:

Location	X (miles)	Y (miles)
Isbelland	220.2	382.1
Pandel	154.5	350.4
Thrunway	360.4	323.2
Joyneria	248.3	287.1
Norvigden	148.7	240.5
Howardan	87.5	300.5
Littmania	63.2	211.7
Bobickstan	111.3	132.6
Essastan	182.9	92.0
Serbankia	206.0	178.6
Balchea	365.3	185.6
Starneria	246.4	49.3

Using the Euclidean distance function as a heuristic, the cost of a node in the below questions is:

$$f(x) = g(x) + h(x)$$

c. What is the total cost of the node ‘Serbankia’ when it is added to the frontier? **(2 points)**

d. What is the total cost of the node ‘Serbankia’ immediately before it is popped from the frontier? **(2 points)**

_____

e. When performing our A* search we find that the node ‘Essastan’ is expanded, and we see our goal, Starneria, in its neighbors. Given that our Euclidean distance function is an admissible heuristic, can we terminate the search immediately upon finding our goal node in this scenario? **(1 point)**

- ◯ Yes
- ◯ No

Provide some *brief* reasoning to your answer:

f. Taking $d(x)$ to be the value of the Euclidean distance heuristic above, which of the following heuristic values are also admissible in the above graph? Select all that apply. **(3 points)**

- ☐ $h(x) = 2 * d(x)$
- ☐ $h(x) = log_2(d(x))$
- ☐ $h(x) = d(x) + 100$
- ☐ $h(x) = d(x) - 100$
- ☐ $h(x) = \frac{1}{2} * d(x)$

3. Optimization Algorithms

(16 points)

Section A.

(9 points)

Genetic algorithms are a unique class of optimization processes popularized by John Holland, a prominent computer scientist who modeled optimization algorithms under the naturally occurring processes of natural selection. Like the name suggests, genetic algorithms utilize the mechanisms found in genetic evolution and apply them to various problem spaces to 'evolve' solutions into even more optimal solutions.

In their search, genetic algorithms process a population of chromosomes (potential solutions), which represents the search space solution, with three operators — selection, crossover, and mutation. The population of individuals go through a sequence of unary (mutation) and higher-order (crossover) transformations. These individuals strive for survival; a selection scheme, biased toward fitter individuals, selects the surviving generations. After some number of generations, the program converges, and the best individual hopefully represents the optimal solution.

Genetic algorithms have had significant success in many modern domains such as encryption, power flow optimization, music composition, and many more. In this section, you'll apply genetic algorithms to a Civil Engineering problem for the design of water distribution systems.

Background:

During the election to her office, the current Mayor of Starneria had promised to revamp the old water pipeline in her city. After assuming office, she learns from her advisors how Genetic Algorithms can help in devising an optimal design for the pipeline. She is challenged by the Isbelland pipe factory that if she finds the most optimal pipe design, they would forego all the operational costs for the next 100 years. Aware that you've learned about Genetic Algorithms in your AI class, she comes to you for help in finding the most optimal network. You start off by trying to understand the problem formulation.

Problem Formulation:

As you can see on the next page, there's a sample pipeline network. There are 4 nodes and 4 links (the pipes). The factory can only produce pipes with diameters of size (in inches) from the set `T = {12, 16, 20, 24}`. Your objective is to find the diameters from the available set for all the links at the minimum possible cost.

An individual 'D' in the population is represented by the set of diameter values for the pipes. The diameter values are arranged in alphabetical order starting with diameter of pipe AB, BC until the last edge.

For example, for **Figure 3.1** (next page), it would be `{AB, BC, CD, DB}`.

Fitness Function:

In the genetic algorithms population, there are some individuals who are 'fitter' than others. To quantify this fitness characteristic, genetic algorithms use what is called a **fitness function** to compare specimens. This function is dependent upon the problem the algorithm is trying to solve, and computes the value we're trying to maximize.

For this particular problem, we'll take advantage of the easily computable cost of the pipeline network. The cost of the network consists of two components - the manufacturing cost of the network and the installation costs. The factory has put out its manufacturing costs for different pipe diameters varying on the length. The manufacturing cost f_{Cost} is given as:

$$f_{Cost} = 1.1 \times L \times D^{1.5}$$

where L is the length of the pipe and D is the diameter in inches and f_{Cost} is the cost in dollars

Thus the manufacturing cost of the network(D) consisting of N pipes becomes

$$f_{Cost}(D) = 1.1 \times \Sigma^{N}_{j} L_j D_j^{1.5}$$

Another aspect to this water distribution problem is the installation cost to handle the water pressure constraints in the network. This term is given by:

$$\gamma(D) = 10000 \times \Sigma_j^N (L_j^{0.5} / D_j^{2.5})$$

And thus, the total cost is given by:

$$C(D) = f_{Cost}(D) + \gamma(D)$$

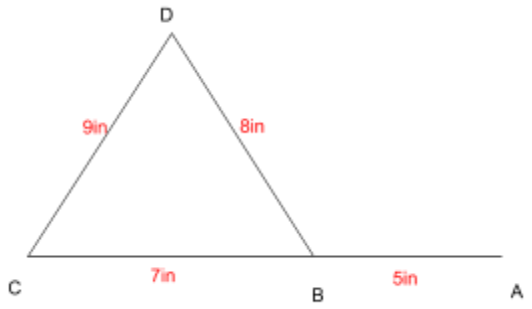


Figure 3.1 : A sample pipeline network

For an individual `S = {16, 20, 12, 24}`,

$$f_{Cost}(S) = 1.1 \times (5 \times 16^{1.5} + 7 \times 20^{1.5} + 9 \times 12^{1.5} + 8 \times 24^{1.5}) = 2486.908676$$

$$\gamma(S) = 10000 \times \left(\frac{5^{0.5}}{16^{2.5}} + \frac{7^{0.5}}{20^{2.5}} + \frac{9^{0.5}}{12^{2.5}} + \frac{8^{0.5}}{24^{2.5}}\right) = 106.790896$$

$$C(S) = f_{Cost}(S) + \gamma(S) = 2593.699572$$

$$\text{Fitness Function} = \frac{1}{C(S)} = 0.000386$$

Now since we want to minimize the cost of the pipeline, we would like a smaller cost network to be more fit than a larger cost one. To account for this, our fitness function for this individual

would be $\frac{1}{C(D)}$, the inverse of the cost. You can see the calculations for the sample network in accordance with the network in **Figure 3.1**.

Now you are given the below pipeline network as shown in **Figure 3.2** for Starneria.

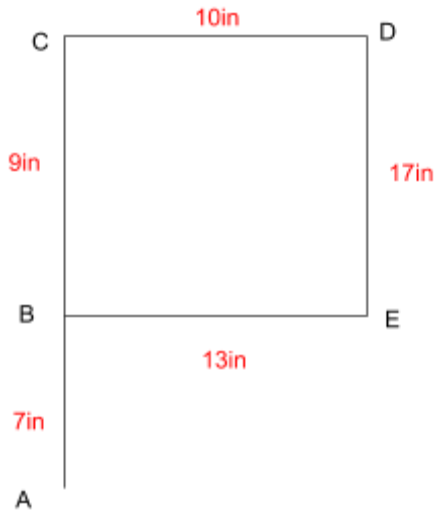


Figure 3.2: Starneria Pipeline Network

Q.3.A.1 (2 points)

Based on the network provided in Figure 3.2, calculate the cost of the given four networks. Remember that the sample {12, 20, 16, 24, 12} represents the diameters for the links A-B, B-C, C-D, D-E and E-B respectively. Report your **final answer** according to the rounding rules of the exam. Please don't round any intermediate value.

Hint: You may find this easier to code.

Specimen	Chromosome	Fitness Score
Network 1	{12, 12, 16, 20, 24}	
Network 2	{24, 24, 12, 12, 16}	
Network 3	{16, 20, 24, 12, 20}	
Network 4	{24, 12, 20, 16, 16}	

Q.3.A.2 (0.5 point)

You're given a budget of $4500 to build this pipeline.

Based on the network provided and your budget which of the following specimens would be feasible? (select all that apply)

- ☐ Network 1
- ☐ Network 2
- ☐ Network 3
- ☐ Network 4

Selection

In order to breed two specimens together to create a new generation, we first have to choose which specimens are going to function as parents. One standard approach to complete this step is called **fitness-proportionate selection** (also referred to as roulette wheel selection). After computing the fitness scores for each specimen in a population, we'll normalize these scores to sum up to one. Using these normalized scores as a probability distribution, we'll select two random parents to pass on their genetic material to the next generation.

Q.3.A.3 (2 points) Using pre-computed costs of the network for the following specimens, calculate the probability that each will be a parent of the next generation. Round your final answer according to the rounding rules of the exam.

Specimen	Fitness Score	Probability to be a Parent
Network 5	0.000316	
Network 6	0.000374	
Network 7	0.000389	
Network 8	0.000347	

Crossover:

After choosing the specimens that will be parents, we have to apply a defined crossover function to generate children from the genetic material of the parents. Unlike the design of the fitness function, there are a number of standard approaches to crossover processes that work sufficiently. For this problem, we use the Arithmetical crossover. That is, the children generated would have genes which are a linear combination of the genes of the parents. Also, an invariant that needs to be maintained during the crossover process is that both the children need to be a valid specimen i.e. in our case, each gene should have a diameter amongst the available diameters. The design for the same is outlined below:

If there are two parents $D_1 = \{d_1^{\,1}, d_2^{\,1}, d_3^{\,1}, d_4^{\,1}, d_5^{\,1}\}$ and $D_2 = \{d_1^{\,2}, d_2^{\,2}, d_3^{\,2}, d_4^{\,2}, d_5^{\,2}\}$ selected for crossover, then the two offsprings J_1 and J_2 generated would be as follows:

$$J_k = \{j_1^{\,k}, j_2^{\,k}, j_3^{\,k}, j_4^{\,k}, j_5^{\,k}\} \text{ for k=1,2}$$

where $j_i^{\,1} = \lambda d_i^{\,1} + (1-\lambda) d_i^{\,2}$ and $j_i^{\,2} = (1-\lambda) d_i^{\,1} + \lambda d_i^{\,2}$ and λ is a constant s.t. $0 \le \lambda \le 1$.

Apart from these rules, we need to do one more check to ensure the children satisfy the constraint that each diameter should belong to the set `T = {12, 16, 20, 24}`

- If $j_i^{\,1}$ does not belong to T, assign the greatest value in T that is less than $j_i^{\,1}$. e.g. if you get $j_i^{\,1}$ as 14, you assign it the value 12.
- If $j_i^{\,2}$ does not belong to T, assign the smallest value in T that is greater than $j_i^{\,2}$. e.g. if you get $j_i^{\,2}$ as 14, you assign it the value 16.

After the crossover step, the children are added to the population for the next generation.

Q.3.A.4 (2 points)

For two parent networks `P1={16, 12, 24, 20, 12}` and `P2={20, 12, 16, 12, 24}`, perform the crossover steps as listed above assuming the value of λ is 0.75. Mark from the below alternatives which would be children of this crossover operation. Select all that apply.

- ☐ {16, 12, 24, 20, 12}
- ☐ {20, 12, 16, 12, 24}
- ☐ {16, 12, 20, 16, 12}
- ☐ {17, 12, 22, 18, 15}
- ☐ {19, 12, 18, 14, 21}
- ☐ {20, 12, 20, 16, 24}
- ☐ {12, 20, 20, 24, 16}
- ☐ {16, 20, 16, 12, 12}

Q.3.A.5 (1 points)

With regards to Arithmetical Crossover, which of the following statements are true? Select all that apply.

- ☐ For $\lambda = 1$, only one new value is added to the next generation's population.
- ☐ For $\lambda = 0.5$, only one new value is added to the next generation's population.
- ☐ For $\lambda = 0.25$, no new values are added to the next generation's population.
- ☐ For $\lambda = 0$, no new values are added to the next generation's population.

Mutation:

In order to introduce new characteristics into a population, the crossover operation is not enough. Again similar to natural populations, we need to 'mutate' children in a small but potentially important manner to traverse our search space efficiently. This method allows children to have characteristics that their parents do not have, and in this way the population as a whole avoids being homogeneous.

In this algorithm, we make use of Gaussian mutation. Given network $D = \{d_1, d_2, ., d_i, .., d_N\}$ as chromosome and gene d_i is randomly selected for mutation, then the gene d_i^* resulting from the application of Gaussian mutation is given by:

$$d_i^* = d_i + 4 \times floor(N(0, \sigma))$$

And the mutated network would be $D^* = \{d_1, d_2, ., d_i^*, .., d_N\}$

Where $N(0, \sigma)$ is an independent random Gaussian number with mean zero and standard deviation $\sigma = 0.1 \times d^{max}$ where d^{max} is the maximum value of the gene and the floor function is the greatest integer function, basically the math.floor method in python.

If the value d_i^* goes out of bound, you simply wrap it around i.e. to get the gene values within the limit, the gene values are adjusted by adding or subtracting multiples of range.

Q.3.A.6 (1.5 points)

Mutate the network D = {16, 16, 24, 12, 20} using the method above. Use $i = 4$ and assume that $N(0, \sigma) = 1.7$. Report the new network generated D^*.

Congratulations! You've completed all the steps of a real-world, relevant application of genetic algorithms and hopefully understand the structure and execution of this class of optimization process. Typically this algorithm is run through many iterations until there's no significant difference in the fitness between generations. The best specimen found at this point is returned as the final result.

Section B.

(7 points)

Alice's Flexible Schedule

Alice is a grad student at Georgia Tech, and she is taking CS6601 this semester. As most of you can relate, her schedule is going crazy because of the workload of the Assignments, and she is trying to find ways to manage her time better.

Let us take a peek at Alice's typical Saturday - the one day each week she gets some time to herself. Out of habit, she always does the following things, in the exact same order.

Please find below the things she does, and the time each thing takes assuming that there is no waiting time.

No.	Task	Time in minutes
1.	She brushes her **teeth** and grabs coffee (T)	30
2.	She goes to the Trailblazer Gym and **works out** (W)	60
3.	She gets back home and takes a **shower** (S)	30
4.	She meets with a friend at **Gibbs' Cafe** for brunch (G)	90
5.	She buys groceries at **A-star Supermarket** (A)	60
6.	She stops to drink **bubble-sort tea** (B)	30

However, she has observed that things don't always go as planned. Some things could take much longer, because of the waiting time. Alas, such is life!

Over the past few weeks, she made the following observations of the waiting time for each task, given the time slot she does it in.

	Waiting time in minutes					
Task	6:00-9:00	9:01-12:00	12:01-15:00	15:01-18:00	18:01-21:00	21:01-00:00
Teeth (T)	30	0	0	0	0	30
Workout (W)	60	30	0	60	30	0
Shower (S)	30	0	0	0	30	0
Gibbs' Cafe (G)	60	30	60	30	120	30
A-star Supermarket (A)	30	90	30	90	30	0
Bubble-sort Tea (B)	0	30	0	60	30	60

(Please keep in mind that waiting depends on the slot in which she **starts** the task. For example, if Alice tries to start working out (W) at 12:00 pm, she will first have to wait 30 mins, causing the entire task will take 90 mins - including waiting time, and so she will finish by 1:30 pm)

Alice, not wanting the things she learned in CS6601 to go to waste, now decides to make use of Simulated Annealing to find out how to minimize the waiting time throughout the course of her day. She wants to pick a time that she should start doing these tasks such that it takes the least amount of overall time to complete. In other words, she wants to do all the tasks while making sure she is left with the maximum amount of time at the end of the day.

NOTE:

- ➢ Alice starts her next task as soon as the previous task is finished. There is no gap between two consecutive tasks
- ➢ All times are in 24-hour format
- ➢ Waiting time for any task outside this range of 6:00 am to 12:00 am can be considered as 0
- ➢ Round all intermediate values to 6 digits, and use those for further computation
- ➢ If the probability of acceptance is greater than 1, it means that the sample is always accepted, and so round it down to 1
- ➢ Total time taken = (Start time - End time), converted to hours. (Note: every half-hour is represented as 0.5, i.e., 1 hour 30 mins would be 1.5 in this column)
- ➢ Energy = $(\text{Total time taken})^{-1}$
- ➢ Temperature = Temperature as used in Simulated Annealing
- ➢ You can assume that every start time is accepted

Help Alice by filling in the table on the next page. Calculate the probability of acceptance for each start time. The first and fourth entries have been filled for you.

Please fill in the following table (2 points):

	Task							Total Time Taken (hours)	Energy	ΔEnergy	Temp-erature	Probability of Acceptance
	T	W	S	G	A	B	End Time					
Start Time of Task	7:00	8:00	10:00	10:30	12:30	14:00	14:30	7.5	0.133333	-	0.07	1
	9:00										0.06	
	15:00										0.05	
	17:00	17:30	19:30	20:30	00:00	01:00	01:30	8.5	0.117647	-0.007353	0.04	0.832081
	5:00										0.03	
	13:00										0.02	
	11:00										0.01	

Given that Alice starts at 9:00, answer the following questions:

Q.3.B.1.a What time does she start taking a shower? **(0.5 points)**

_____

Q.3.B.1.b What time does she finish all her tasks?**(0.5 points)**

_____

Q.3.B.1.c How many hours did she take to do all her tasks?**(0.5 points)**

_____

Q.3.B.1.d What is ΔEnergy for this start time?**(0.5 points)**

_____

Q.3.B.1.e What is the probability of acceptance for this start time?**(0.5 points)**

_____

Given that Alice starts at 11:00, answer the following questions:

Q.3.B.2a What time does she finish buying groceries from A-Star Supermarket?**(0.5 ps)**

_____

Q.3.B.2b What time does she finish all her tasks? **(0.5 points)**

_____

Q.3.B.2c How many hours did she take to do all her tasks? **(0.5 points)**

_____

Q.3.B.2d What is ΔEnergy for this start time? **(0.5 points)**

_____

Q.3.B.2e What is the probability of acceptance for this start time? **(0.5 points)**

_____

4. Constraint Satisfaction Problems

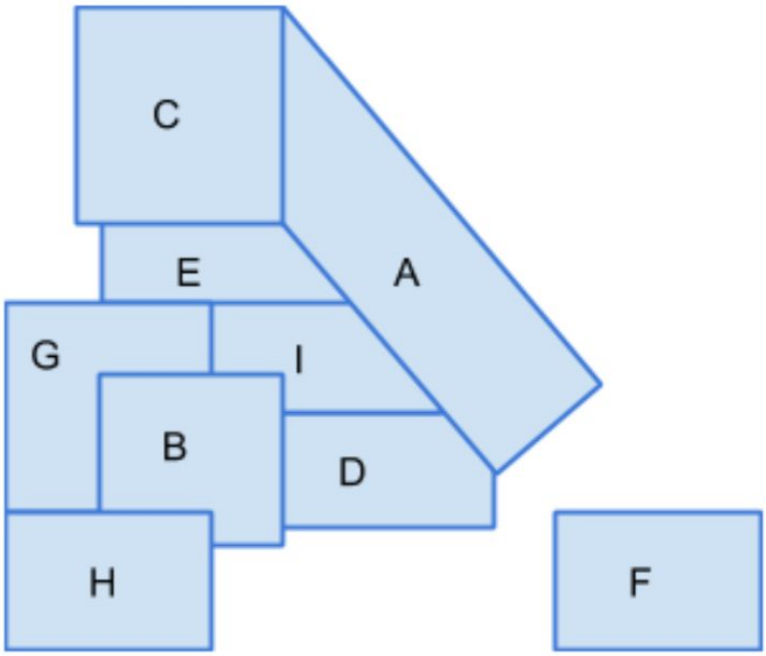
(16 points)

Question A.

(9 points)

The year is 6601 A.D. and after years of conflict, including pranking and petty bickering, the nations of Starneria and Isbelland have finally settled their differences. Now that the citizens of each nation are free to travel back and forth, you, an AI-savvy entrepreneur, decide to start a travel agency that claims to craft itineraries for clients no matter their constraints. Since Isbellians have not, until now, been very familiar with the best places to visit within Starneria, there are many Isbellians that are expecting you, with all of your AI knowledge in constraint-satisfaction, to be an expert that will help them plan their awaited vacations to Starneria.

The wonderful land of Starneria is comprised of 9 regions: A, B, C, D, E, F, G, H, and I. Each region boasts its own popular sites and attractions, but not every region will be right for everyone.



Region A - Contains the capital of Starneria located in the north, full of museums, plays, and shopping areas
Region B - Southern region, known for being notably hot all year. It has a small stretch of coastline where much of the nation's goods are imported at a major port.
Region C - Mountainous region in the north, popular among hikers
Region D - A coastal region in the south with its share of beaches. It is popular even among Starnarians, resulting in many shops and restaurants.
Region E - A northern region that receives a noteworthy amount of rain. Much of the nation's crops are grown here.
Region F - A southern island with beautiful beaches and coastline that is only accessible by boat from Region B or D, as there is no domestic airport in this region.
Region G - While technically a central region, this is officially considered part of the north. Many cultural sites are present here including museums. Shopping boutiques are always nearby.
Region H - Desert region in the south, with the hottest temperatures in Starneria, but also home to the largest markets in the nation. Any shopper would love visiting this region.
Region I - Home of the only international airport in the country. With its central location, it offers the best of the northern and southern cultures. Still, it is officially part of the north as it shares more in common with the other northern regions.

Note#1 - Any region can be accessed from any other region (via domestic flights) **except** for Region F which must be entered/exited via boat from/to Region B or Region D.
Note#2 - All tourists visiting Starneria must enter and exit via an international airport for legal reasons.

Jack, Jane, John, and Jamie come to you for help with planning their itinerary. Each member of the group has different things they want out of the trip, but each also has restrictions to work around. Each person doesn't care where they go on the trip as long as their personal constraints are met.

Jack - Considering that Isbelland has no coastal borders, he wants to spend time by the water in at least 2 coastal regions. In particular, he says he must go to an island in order to have a good trip.
Jane - Due to health conditions, she is unable to travel to regions with high heat and is unable to travel to C or E due to allergies. She'd really like to go to Region G to see a specific museum during the trip.
John - He wants to do some shopping for some great Starnerian souvenirs, but after his involvement in The Great Prank of 6586 he is still banned from entering regions A and D.
Jamie - She wants to see as much as possible, and is confident that she will be happy in any of the regions. Since she heard the north and south are very different in Starnaria, she definitely wants to make sure she sees at least 2 northern regions as well as at least 2 southern regions.

Q.4.A.1 Check all clients with unary constraints that will affect the group's itinerary **(1 point)**

- ☐ Jack
- ☐ Jane
- ☐ John
- ☐ Jamie
- ☐ None

Q.4.A.2 In the following table, place an "X" on all regions that each client would want to visit. Also place an "O" on all regions that the client is open to visiting. That is, place an "O" if the client has no opposition toward going to the corresponding region. Leave cells blank where a client is not able to visit the corresponding regions. If multiple regions can be used to satisfy a constraint, place an X on all that apply. **(3 points)**

Since Jane wants to go to Region G, the first X has been placed for you to reflect this.

	A	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**
Jack									
Jane							X		
John									
Jamie									

Q.4.A.3.a In order for an itinerary to be a valid option, all group members must be able to visit all regions on the itinerary together. Based on the given constraints, is it possible for you to create an itinerary that works for this entire group? **(0.5 points)**

- ◯ Yes
- ◯ No

Q.4.A.3.b If yes, write the sequence of regions on the itinerary with the minimum number of regions visited. Otherwise, just list the regions that the group can visit together. Break ties alphabetically (for example, if you could choose either A or B, choose A first). **(2 points)**

_____

Just before you are about to report back to the group, you learn that Jane will be able to use a new medication that will allow her to enter regions with high heat. Considering this new information, you re-evaluate the options.

Q.4.A.4.a With this new information, is it possible for you to create an itinerary that works for this entire group? **(0.5 points)**

- ◯ Yes
- ◯ No

Q.4.A.4.b If yes, write the sequence of regions on the itinerary with the minimum number of regions visited. Otherwise, just list the regions that the group can visit together. Break ties alphabetically (for example, if you could choose either A or B, choose A first). **(2 points)**

_____

Question B.

(7 points)

Famished from their travels across Starneria, Jack, Jane, John, and Jamie take a break from their adventure to enjoy some of Starneria's national dish: pizza.

The travelers have the following topping preferences and **they each want two slices of pizza**:

Jamie: Vegetarian. Standard marinara sauce. No meat containing slices can be directly adjacent.
John: Wants slices with pineapple. Standard marinara sauce, no adjacent slices can have alfredo sauce.
Jack: Wanting to get the full Starneria experience, only wants to eat Professor Starner's favorite: chicken with alfredo sauce.
Jane: Still an Isbelland hold out, she refuses to eat alfredo sauce and wants italian sausage as a topping.

Since the Starnerian civilization is one of the utmost technically developed, the pizza is made by a robot. The robot, however, has two limitations:

1. To avoid contamination, the machine places non-meat toppings (vegetables & fruits) first and then switches to meat toppings. After switching to meat, it cannot go back to vegetable toppings.
2. To keep costs to a minimum, the machine can only switch sauces once.

For practical considerations, the variables will be encoded as the following:

V = vegetarian
P = pineapple
C = chicken alfredo
S = italian sausage
M = marinara sauce
A = alfredo sauce

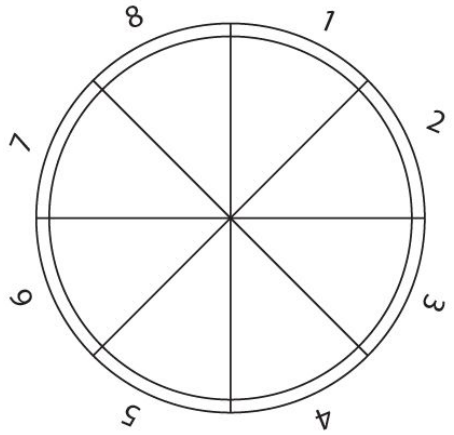
Q.4.B.1 Which of the constraints is unary? **(1.5 points)**

- ◯ No marinara sauce
- ◯ No adjacent slice can be meat
- ◯ No alfredo sauce
- ◯ The machine starts with vegetarian toppings

Q.4.B.2 Which of the following constraints are binary? **(1.5 points)**

- ◯ No marinara sauce
- ◯ No adjacent slice can be meat
- ◯ No alfredo sauce
- ◯ The machine starts with vegetarian toppings

Q.4.B.3 Using the minimum remaining values heuristic, solve the pizza topping CSP problem. Fill the table below accordingly. **(4 points)**



Slice Number	Topping	Sauce
1		
2		
3		
4		
5		
6		
7		
8		

5. Probability

(16 points)

Section A

(7 points)

In a Starnerian medical school, a doctor gives a patient a test for particular cancer, but before the results of the test, the only evidence the doctor has is that 1 in 10000 patients have this cancer. Additionally, experimental research at the school has shown that, in 99 percent of the cases in which cancer is present, the test is positive; and in 95 percent of the cases in which it is not present, it is negative.

Q.5.A.1 If the test turns out to be positive, what probability should the doctor assign to the event that cancer is present?

_____

Q.5.A.2 Why is the probability calculated in the previous question so low (less than 0.5) despite the fact that 99 percent of cases in which cancer was present the test was positive?

- ☐ The answer is expected and not lower than 0.5.
- ☐ Because the prior probability of having cancer is much larger than the probability of the test being positive.
- ☐ Because the prior probability of having cancer is much smaller than the probability of the test being positive.
- ☐ Because we did not take into account the normalization constant needed to make the probability sum to 1 for `P(cancer | positive)`.

Q.5.A.3 Among the results where the test is positive, what is the false-positive rate (i.e., what is the percentage of positives that are due to error)?

_____

Q.5.A.4 What must be the prior probability of cancer, to result in a posterior probability of 0.5 of having cancer given a positive test result?

Q.5.A.5 Now suppose from another study doctors have found out that in addition to the test being positive or negative, the patient's gender is also a factor that contributes to the probability of having cancer. With this knowledge, the doctor needs to formulate the probability the patient has cancer given the test is positive and gender is male (i.e., $P(Cancer \mid test = positive,\ gender = male)$). As a probability expert, which of the following statements do you think are true (select all that apply):

☐ You can use Bayes rule to calculate this probability as:

$$P(cancer \mid test = P\ ,\ gender = M) = C \cdot P(test = P\ ,\ gender = M \mid cancer) \cdot P(cancer)\ .$$

☐ You cannot use Bayes rule because we have 2 dependent variables to take into account.

☐ Using conditional independence assumption we can calculate the probability as:

$$P(cancer \mid test = P\ ,\ gender = M) = C * P(test = P \mid cancer) * P(gender = M \mid cancer) * P(Cancer)$$

☐ You can calculate this probability as:

$$P(cancer \mid test = P\ ,\ gender = M) = C * P(test = P\ ) * P(gender = M) * P(Cancer)$$

Note: *C* is a normalization constant, P is Positive test, M is gender male.

Section B

(9 points)

Game 1 - Roulette

Consider a game of Roulette. There are 18 black slots, 19 red slots, and 1 green slot. A person wants to gamble and has a 40% chance of betting on black, 50% chance of betting on red and a 10% chance of betting on green. The ball has an equal probability of landing into any of the slots. If the ball lands into the same color slot like the one on which the bet was placed, then the person wins.

Q.5.B.1.1 What is the probability of winning? **(0.5 point)**

_____

Q.5.B.1.2 Suppose that the person loses the round. What is the probability that he bet on red? **(1 point)**

_____

Q.5.B.1.3 If you place a bet on red or black and win, your money is doubled, and if you place your bet on green and win, then your money is tripled. Suppose you bet $1000 in a round of roulette. What is the expected amount that you will have after the round? **(1 point)**

_____

Q.5.B.1.4 What betting strategy should you follow so that on average you do not lose any money while playing the game in **Q.5.B.1.3**? **(1.5 points)**

a. Probability of betting on Black: _____

b. Probability of betting on Red: _____

c. Probability of betting on Green: _____

Game 2 - Double or Nothing

Charles Isbell has been running short of cash lately. To earn more money he decides to start a game called “Double or Nothing”. A gambler starts with a fortune of $1000. The game involves multiple rounds of gambling. If they win a round, they gain $100, otherwise, they lose $100. The game ends only when the gambler loses all of their money (reaches $0) or reaches the goal of $2000. The probability of winning a single round is x and is independent of the past rounds. Charles has challenged Thad to play a game of Double or Nothing. Thad is a smart guy and wants to analyze his chances of winning before accepting the challenge.

Q.5.B.2.1 Let P_r be the probability that a gambler starting with a fortune of $\$r$ will reach the goal of $2000 eventually and win the game. For example, $P_0 = 0$ because a gambler with a fortune of $0 can never reach $2000 and $P_{2000} = 1$ because he has already reached the goal. P_i can be written in terms of P_{i+100} and P_{i-100} depending on whether the gambler wins a round or not. Let us say $P_i = aP_{i+100} + bP_{i-100}$ for $i \in \{100, 200, \ldots 1900\}$. What are the values of a and b (in terms of x)? **(1 point)**

a = _____

b = _____

Q.5.B.2.2 Let's say that $x = 0.5$. What is the probability that Thad (the gambler) will win the game? In other words, what is P_{1000}? **(3 points)**

Hint: We can simplify the recurrence relation in the previous part using $P_i = xP_i + (1-x)P_i$. After rearranging the terms, we get something like:

$$P_{i+100} - P_i = R(P_i - P_{i-100}) = R(R(P_{i-100} - P_{i-200})) = \ldots = R^{i/100}(P_{100} - P_0)$$

, where R is a function of x.

Adding the terms $P_{i+100} - P_i$ for $i \in \{0, 100, \ldots, 1900\}$ will be helpful for getting P_{100} which can be used while solving for the final answer.

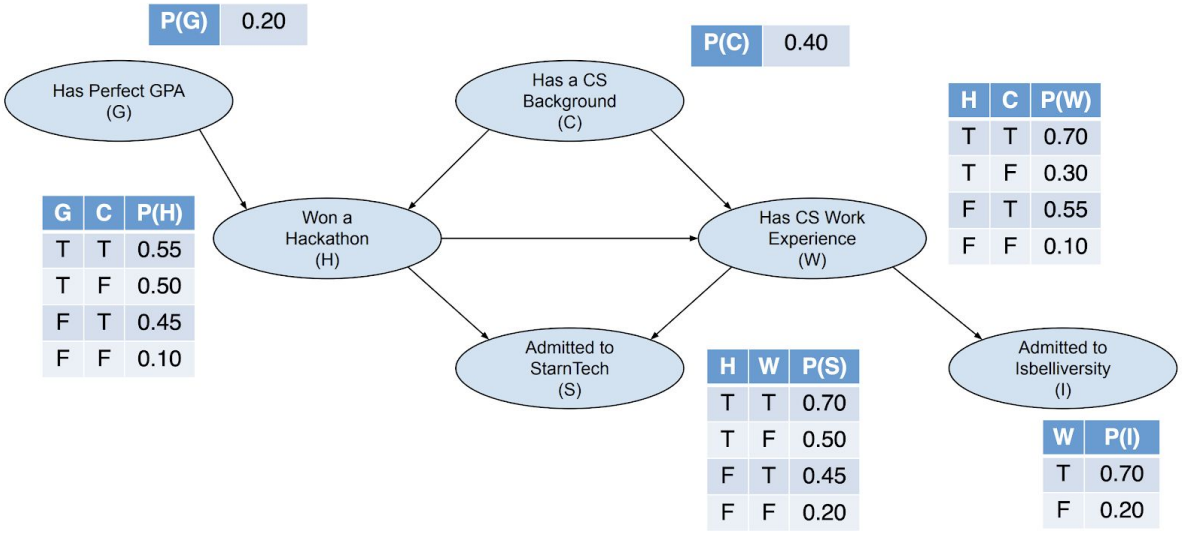
Q.5.B.2.3 Charles realized that he had not been making much money out of this game. So, he changes the probability of winning a single round to $x = 0.45$ without telling anyone. What is the new probability that Thad will win the game? **(1 point)**

Hint: For $R \neq 1$, $1 + R + R^2 + R^3 + \ldots + R^n = (1 - R^{n+1})/(1-R)$

6. Bayes Nets

(16 points)

StarnTech and Isbelliversity have been in college rivalry with each other over the years. This rivalry is not only in academics and athletics but also in the rigorous admissions process to build a stronger incoming cohort. The Computer Science admissions committee from both universities have recently made public the factors that will be evaluated in order to grant admissions to students in their Master's Program. The following Bayesian Network and the corresponding conditional probability tables describe the admission criteria used.



P(G)	0.20

P(C)	0.40

G	C	P(H)
T	T	0.55
T	F	0.50
F	T	0.45
F	F	0.10

H	C	P(W)
T	T	0.70
T	F	0.30
F	T	0.55
F	F	0.10

H	W	P(S)
T	T	0.70
T	F	0.50
F	T	0.45
F	F	0.20

W	P(I)
T	0.70
F	0.20

Part A

D-Separation and Conceptual Questions (7 points)

Given the above Bayes Network, answer the following questions.

Q.6.A.1 The above Bayes Network is: **(0.5 points)**

- ◯ Singly connected
- ◯ Multiply connected

Q.6.A.2 The time and space complexity of Variable Elimination in the worst-case scenario for the above Bayes Net is: **(0.5 points)**

- ◯ Linear in the size of the network
- ◯ Exponential in the size of the network
- ◯ None of the above

Q.6.A.3 Consider the claim: "G is conditionally independent of S, given X". For which of the following values of X is the claim true? Choose all such values below. **(1.5 points)**

- [ ] C
- [ ] H
- [ ] W
- [ ] I
- [ ] H, W

Q.6.A.4 Consider the claim: "S is conditionally independent of I, given X". For which of the following values of X is the claim true? Choose all such values below. **(1.5 points)**

- [ ] G
- [ ] C
- [ ] H
- [ ] W
- [ ] H, W

Q.6.A.5 Consider the claim: "G is conditionally independent of I, given X". For which of the following values of X is the claim true? Choose all such values below. **(1.5 points)**

- [ ] C
- [ ] H
- [ ] W
- [ ] S
- [ ] C, S

Q.6.A.5 Assume **for this question only** that the edge H → W is removed. For the resulting Bayes Network, select all nodes that belong to the Markov Blanket of W: **(1.5 points)**

- [ ] G
- [ ] C
- [ ] H
- [ ] S
- [ ] I

Part B.

Exact Inference (9 points)

With the provided Bayesian Network, calculate the following probabilities. Please do not round off the *intermediate* probability values. For the following questions, we will require you to *show your work* to demonstrate how you computed the various probabilities.

Q.6.B.1 Compute the probability of John getting rejected by Isbelliversity given that he has no work experience and has never won a hackathon, but has a perfect GPA. **(2 points)**

_____

Show your work:

Q.6.B.2.a Compute the probability of Rosa getting admitted to StarnTech given that she does not have a CS background but has won a hackathon. Use Variable Elimination. **(2 points)**

Q.6.B.2.b Select the variables over which the "sum out" operation is performed in the above question. **(1 point)**

- ☐ G
- ☐ C
- ☐ H
- ☐ W
- ☐ S
- ☐ I

Show your work:

Q.6.B.3 It is known that the best university is the one that has a lower probability of getting admission.

Q.6.B.3.a Which University is best? **(1 point)**

- ◯ Isbelliversity
- ◯ StarnTech
- ◯ Both are the same

Q.6.B.3.b What is the difference between the two probabilities? **(3 points)**

_____

Show your work:

7. Machine Learning

(16 points)

Part 1

It is election day at GatechLand and students all across the country voted to decide the runoff between this year's presidential candidates - Thadonix and Isbellatrix. A group of students decided to predict this year's state-wise wins. They used the KNN (K-nearest neighbor) algorithm with Euclidean distance on 3 state features <AI Research, Pizza Availability, Climate Change Crusader> to predict the state majority. The model was trained on the below dataset (10 labeled data points where each row represents a single data point):

State	AI Research	Pizza Availability	Climate Change Crusader	Majority Win
MS	9	21	7	Thad
VB	6	12	10	Thad
MK	5	5	9	Thad
KK	2	-8	-7	Isbell
BV	16	-3	-7	Isbell
BB	-6	-6	-8	Thad
GP	8	6	9	Isbell
CR	3	-3	-8	Isbell
SV	-10	10	8	Thad
AA	10	-8	7	Isbell

Q.7.1.a Based on the above dataset fill in the below table by computing the 3 nearest neighbors for each of the given states. **(3 points)**

State	1	2	3
VB			
BB			
SV			
AA			

Q.7.1.b Fill in the table by computing the leave-one-out cross-validation error (expressed as a number of misclassifications) for each of the k values. **(3 points)**

K	Error
3	
5	
7	

Q.7.1.c If you can change the value of 'K' in KNN, what is the **least value of K** for which you are certain of state AA having a Thadonix majority win? **(1 points)**

Q.7.1.d If we train the above dataset on a 1NN classifier, then we would achieve a 100% training accuracy. **(1 points)**

- ◯ True
- ◯ False

Q.7.1.e Which family of algorithms does this classifier belong to? **(1 points)**

- ◯ Supervised learning
- ◯ Unsupervised learning
- ◯ Semi-supervised learning

Part 2

Vicky is a graduate student at StarnTech and she plans to spend her Spring Break by binge-watching the top 5 highly rated Marvel Cinematic Universe movies. (She does not trust the IMDb or Rotten Tomatoes ratings) In order to help her select these movies, she builds a Logistic Regression model for predicting the movie ratings based on Google trends. She believes that she has built a fairly accurate model and she wishes to evaluate this against the ratings given by her friend. She defines the below evaluation metric to carry out this task:

a. Highly rated: Ratings > 3.5
b. Poorly rated: Ratings ≤ 3.5

NOTE: The friend's rating is treated as the ground truth for the below evaluation.

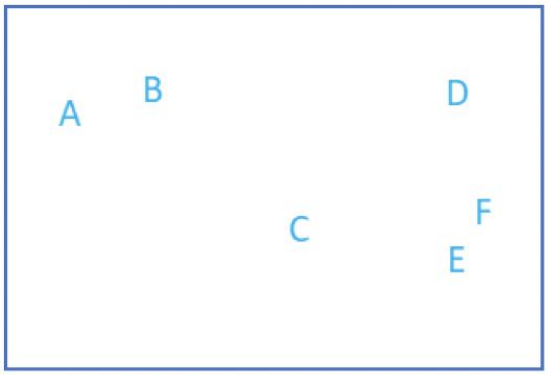
Movie	**Vicky's model predictions**	**Friend's rating**
Avengers: Infinity War	4.6	4.3
Captain Marvel	3.9	3.4
Black Panther	3.2	4.1
Avengers: Endgame	4.0	4.8
Thor Ragnarok	4.7	4.4
Ant-Man	3.6	2.9
Guardians of the Galaxy	3.0	3.7
Captain America: The First Avenger	4.5	3.2
Iron Man	4.1	4.7

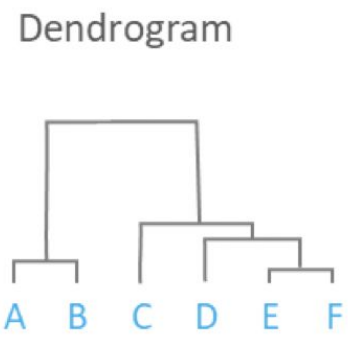
Q.7.2 Based on the above data calculate the below evaluation metrics on the success of the system for correctly rating movies as either highly rated or poorly rated: **(2 points)**

Precision: _____ Recall: _____

Part 3

A popular clustering method not focused on in this class is called **hierarchical clustering.** In this algorithm, the goal is to form a hierarchy of clusters in your data. Unlike K-Means and Expectation-Maximization, we don't need to assign a value to the number of clusters we're looking for; hierarchical clustering will perform clustering with different numbers of clusters to see how some clusters can decompose into others. To clarify, imagine we have a dataset with 6 points that we want to cluster using this method. The dataset and clustering output would look like the following:





Looking at the dendrogram on the right, we can see that when the number of clusters is 2, the clusters are [A,B] and [C,D,E,F]. When the number of clusters is 3, we have [A,B], [C], [D,E,F]. In this way, we can see how clusters decompose into each other, and how clusters within our dataset relate to each other.

An outline of one implementation of this algorithm to build this dendrogram bottom-up is shown below:

1. Initialize each data point in your dataset to be its own cluster.
2. Choose an appropriate distance metric
3. Until you have 1 cluster:
    a. Calculate the minimum distance from every cluster to every other cluster by the chosen distance metric.
    b. Merge the two closest clusters

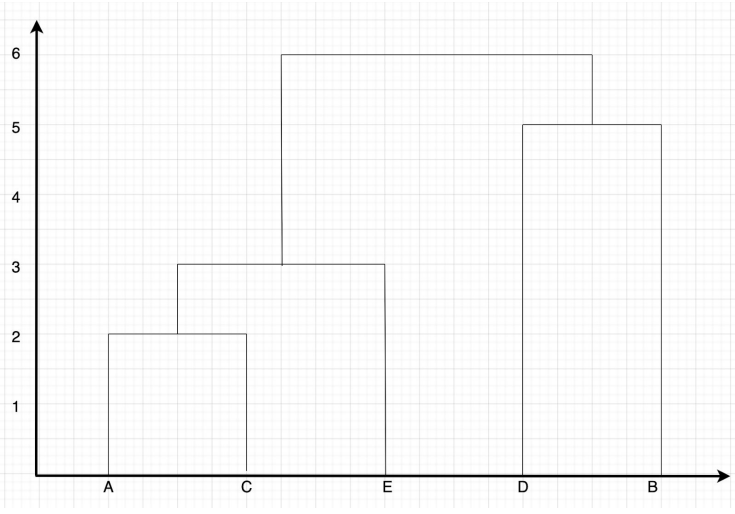
Some of the most commonly used distance metrics between the two clusters are:

1. **Single linkage:** the minimum distance between any pairs of points from the two clusters
2. **Complete linkage:** the maximum distance between any parts of points from the two clusters
3. **Average linkage:** the average distance between all pair of points from the two clusters

	A	**B**	**C**	**D**	**E**
A	0				
B	9	0			
C	2	7	0		
D	6	5	9	0	
E	11	10	3	8	0

To help you get started we have completed the dendrogram and the iteration table using single linkage.

Iteration	**Cluster Representation**
Iteration 0	[A], [B], [C], [D], [E]
Iteration 1	[A,C], [B], [D], [E]
Iteration 2	[A,C,E], [B], [D]
Iteration 3	[A,C,E], [D,B]
Iteration 4	[A,C,E,D,B]



Q.7.3.a With this understanding of the hierarchical clustering algorithm and the above distance matrix fill in the iteration table as shown above for complete linkage distance metric. **(3 points)**

Iteration 0	[A], [B], [C], [D], [E]
Iteration 1	
Iteration 2	
Iteration 3	
Iteration 4	[A,B,C,D,E]

Q.7.3.b When using average linkage distance metric which of the below options represent cluster formations when the number of clusters is 2. **(2 points)**

- ◯ [A,C,E], [B, D]
- ◯ [B, C, D], [A, E]
- ◯ [A, C, B], [D, E]

Checklist

Mark the checklist below making sure you have taken care of each of the points mentioned:

- ☐ I have read the pinned Piazza post with the title 'Midterm Exam Clarifications Thread', and I am familiar with all of the clarifications made by the Teaching staff.
- ☐ All answers with more than 6 digits after the decimal point have been rounded to 6 decimal places.
- ☐ All pages are being uploaded in the correct order that they were presented to me.
- ☐ Any extra pages **(including blanks)** are only attached at the END of this exam, after page 46 with clear pointers to wherever the actual answer is in the PDF (reference properly).
- ☐ I am submitting only one PDF and nothing else (no docx, doc, etc.).
- ☐ The PDF I am submitting is not blank (unless I want it to be)
- ☐ I will upload a PDF on both Gradescope and Canvas (for backup).
- ☐ I have done my own work and not collaborated with anyone else nor used materials besides those allowed in the instructions at the beginning of this exam.
- ☐ **I will go over my uploads (especially any pictures) on Gradescope and make sure that all the answers are clearly visible. I acknowledge that I am aware that dull / illegible / uneven scans and submission that don't follow the above guidelines will not be graded.**