Assignment 0

CSE 447 and 517: Natural Language Processing - University of Washington

Winter 2022

Please consult the course website for current information on the due date, the late policy, and any data you need to download for this assignment.

This assignment is intended as a tool for students who have an interest in taking CSE 447 or CSE 517 (Natural Language Processing) but are not enrolled in the programs these courses were designed for. You should complete it on your own before committing to take the class. **You're welcome to use any textbook or online references when completing this assignment.** The goal is not to check whether you remember all the detailed information required to get to a solution, but rather to see that you have internalized enough mathematical background to know, roughly, how to proceed, and how to find the details. Note that this assignment is neither an exhaustive "checklist" of everything you might need to know before starting this course, nor will all of the specific concepts here necessarily make an appearance in the course. The assignment is an approximation.

The assignment covers:

- Probability and statistics
- Linear algebra
- Calculus
- Dynamic programming
- Reflections on natural language

This assignment is not graded. You should evaluate your own answers using the solutions given. Instead of thinking of this as a "pass or fail" assignment, you should use it as an estimate of the amount of extra time you might need to put in to achieve your own goals for this class. For example:

- If you are comfortable doing every problem and scored very well (say, all but one of the problems mostly correct), then you shouldn't have too many problems with the mathematical content of CSE 447/517.
- If you found that you didn't know how to get started with many of the problems, and the concepts used in the solutions were unfamiliar to you, then you might not be ready for CSE 447/517 yet.[1]

[1]Recommended courses for probability & statistics: CSE 312, STAT 390/391; for linear algebra: MATH 308, MATH 318; for calculus: MATH 126.

- If you found the challenge somewhere between those two points (e.g., you needed to review material from courses you took in the past or fill in some gaps using online materials), then you should expect to need extra time in CSE 447/517 if you want to do well. Every year some students in this situation take the course, and some do quite well; you should enroll only if you have the extra time and energy to spend on the course.

When we can't think for ourselves,
we can always quote.

LUDWIG WITTGENSTEIN

1 Probability and Statistics

Mango

Suppose there are two bags. Bag 1 contains 4 mangoes and 2 apples. Bag 2 contains 4 mangoes, 2 apples, and 2 bananas. There is also a biased coin that comes up heads with probability 0.6 (and tails with probability 0.4).

Your friend goes behind a screen to toss the coin, and chooses a fruit (uniformly at random) from bag 1 if it comes up heads, or from bag 2 if it comes up tails. Your friend comes out from behind the screen and hands you a mango; you don't know which way the coin landed or which bag the mango came from.

What is the probability that the mango was picked from bag 2?

Running

Consider a sequence of coin tosses, which we can encode as a binary string (1 will mean heads, 0 will mean tails). Let a "run of n tails" mean a sequence of (exactly) n tails, preceded by a heads or the beginning of the sequence, and followed by a heads or the end of the sequence. For example, in this encoding of a sequence of coin tosses, there are two runs of 3 tails: 11000101110001. If you toss a fair coin 100 times, what is the expected number of runs of six tails?

Bias Detection

Your friend has two coins; coin A is a fair coin worth \$1,000 and coin B is biased and worth \$1,200 because its bias was induced by a famous NLP researcher. When tossed, the fair coin (A) comes up heads (1) with probability $\frac{1}{2}$. The biased coin (B) comes up tails (0) with probability $\frac{3}{4}$. Your friend chooses one of the coins at random (equal probability of choosing coin A or coin B); you want to determine which coin it is, using statistical reasoning. You toss the coin 5 times independently and observe $(0, 0, 1, 0, 1)$ before your friend snatches it back. What is the probability that the coin you were given was coin B, given what you have observed?

Suppose your friend offers to give you the coin to keep, but only if you can correctly guess whether it was fair coin A or biased coin B. What is your guess and why?

2 Linear Algebra

Random Diagonal

Let $\mathbf{A} = \begin{bmatrix} a_{1,1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{n,n} \end{bmatrix}$, where all entries of $\mathbf{A}$ are 0 except its diagonal entries $(a_{1,1}, \dots, a_{n,n})$. Now suppose that each diagonal entry $a_{i,i}$ is drawn uniformly from the range $[-1, 1]$ (consider it random variable $A_{i,i}$), and assume that $n > 1$. Find:

- $p(A_{1,1} = 0)$
- $p(A_{n,n} > 0.5 \mid A_{n-1,n-1} \geq 0)$
- $p(\text{rank}(\mathbf{A}) < n)$
- $p(\mathbf{A} \succeq 0)$, i.e., the probability $\mathbf{A}$ is positive semidefinite (hint: if $n = 2$, what are the eigenvalues of $\mathbf{A}$?)

Matrix Operations

Let $\mathbf{B} = \begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}$.

- Is $\mathbf{B}$ invertible? If so, find $\mathbf{B}^{-1}$.
- Is $\mathbf{B}$ diagonalizable? If so, find its diagonalization.

3 Calculus

Derivatives of Activation Functions

The "sigmoid" and hyperbolic tangent functions are commonly used in neural networks to monotonically map a real-valued scalar into a finite range.

- Let $\sigma(x) = \frac{1}{1+e^{-x}}$. Write $\frac{d\sigma}{dx}$ in terms of $\sigma(x)$.
- Let $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. Write $\frac{d}{dx}\tanh(x)$ in terms of σ and x.

4 Dynamic Programming

Tile Collection

Consider an $n \times n$ grid. At position (i, j) in the grid, there is a reward $r_{i,j} > 0$. You want to travel from the top left tile (at $i = j = 1$) to the bottom right tile ($i = j = n$) while collecting rewards on the tiles you've visited. You can only move down or right one tile in each step (not diagonally).

Design a dynamic programming algorithm that maximizes (and outputs) the sum of your reward when you reach the bottom right tile. A 3×3 example is shown below; the optimal path is colored in red. Give the time and space complexity of your algorithm.

1	2	8
6	5	5
3	4	1

5 Reflections on Natural Language

Lamp and Box

Consider the two sentences:

1. The lamp won't fit in the box because it is too big.
2. The lamp won't fit in the box because it is too small.

For each sentence, read it, then ask yourself what *it* refers to. What is your answer for sentence 1, and what is your answer for sentence 2? Can you create another sentence pair like these, where only one word is different, but the meaning of another word (a pronoun) changes? What do you think a computer program would require to make such judgments the same way a human typically would?

Solutions

Mango There are three events. First, the coin toss, random variable C, which ranges over $\{\text{heads}, \text{tails}\}$, and you know that $p(C = \text{heads}) = 0.6$. Next, the bag, B, which ranges over $\{1, 2\}$. Last, the fruit, F, which ranges over $\{\text{apple}, \text{banana}, \text{mango}\}$. You are interested in $p(B = 2 \mid F = \text{mango})$. By the definition of conditional probability,

$$p(B = 2 \mid F = \text{mango}) = \frac{p(B = 2, F = \text{mango})}{p(F = \text{mango})}. \tag{1}$$

The terms in both the numerator and denominator fail to mention random variable C, and the denominator also fails to mention random variable B. We must therefore rewrite each as a marginal probability that sums over all possible values of the unmentioned random variables:

$$= \frac{\sum_{c \in \{\text{heads,tails}\}} p(C = c, B = 2, F = \text{mango})}{\sum_{c \in \{\text{heads,tails}\}} \sum_{b \in \{1,2\}} p(C = c, B = b, F = \text{mango})} \tag{2}$$

These joint probabilities can be factored using the chain rule. We'll order the random variables as C, then B, then F, to fit the story:

$$= \frac{\sum_{c \in \{\text{heads,tails}\}} p(C = c) \cdot p(B = 2 \mid C = c) \cdot p(F = \text{mango} \mid C = c, B = 2)}{\sum_{c \in \{\text{heads,tails}\}} \sum_{b \in \{1,2\}} p(C = c) \cdot p(B = b \mid C = c) \cdot p(F = \text{mango} \mid C = c, B = b)} \tag{3}$$

We can drop C from the final factor in both the top and the bottom, because once the bag is chosen, the coin toss doesn't matter (there's conditional independence between the fruit and the coin, given the bag):

$$= \frac{\sum_{c \in \{\text{heads,tails}\}} p(C = c) \cdot p(B = 2 \mid C = c) \cdot p(F = \text{mango} \mid B = 2)}{\sum_{c \in \{\text{heads,tails}\}} \sum_{b \in \{1,2\}} p(C = c) \cdot p(B = b \mid C = c) \cdot p(F = \text{mango} \mid B = b)} \tag{4}$$

Let's unfold the numerator and plug in the values provided by the problem:

$$p(C = \text{heads}) \cdot p(B = 2 \mid C = \text{heads}) \cdot p(F = \text{mango} \mid B = 2) \tag{5}$$
$$+ p(C = \text{tails}) \cdot p(B = 2 \mid C = \text{tails}) \cdot p(F = \text{mango} \mid B = 2) \tag{6}$$
$$= 0.6 \cdot 0 \cdot \frac{4}{8} + 0.4 \cdot 1 \cdot \frac{4}{8} = 0.2 \tag{7}$$

Now we unfold the denominator, which has four terms. Two of them are zero and are not included here, for clarity.

$$p(C = \text{heads}) \cdot p(B = 1 \mid C = \text{heads}) \cdot p(F = \text{mango} \mid B = 1) \tag{8}$$
$$+ p(C = \text{tails}) \cdot p(B = 2 \mid C = \text{tails}) \cdot p(F = \text{mango} \mid B = 2) \tag{9}$$
$$= 0.6 \cdot 1 \cdot \frac{4}{6} + 0.4 \cdot 1 \cdot \frac{4}{8} = 0.4 + 0.2 = 0.6 \tag{10}$$

The final answer, then, is $\frac{0.2}{0.6} = \frac{1}{3}$.

Running The quantity we are interested in is $\mathbb{E}[\text{instances of runs of six tails}]$ which isn't very mathematical. Let the 100 binary random variables be denoted $X_1, \ldots, X_{100}$; the value 0 will denote tails and 1 heads. Let X_0 and X_{101} denote the non-coin tosses at the beginning and end of the sequence, respectively, which can never take the value 0 (or 1). The key is to use linearity of expectation to break the desired quantity down into a sum of 95 much easier expectations:

$$\mathbb{E}[\text{instances of runs of six tails}] = \sum_{i=1}^{95} \mathbb{E}[1 \text{ if a run starts at position } i, 0 \text{ otherwise}] \tag{11}$$

$$= \sum_{i=1}^{95} p\left(\begin{array}{c} X_{i-1} \neq 0, X_i = 0, X_{i+1} = 0, X_{i+2} = 0, \\ X_{i+3} = 0, X_{i+4} = 0, X_{i+5} = 0, X_{i+6} \neq 0 \end{array} \right) \tag{12}$$

We don't worry about positions 96 and above, because a run of six can't start there (X_{101} can't be 0); the last position for one of our runs to start is position 95. You may have an uneasy feeling about this transformation, because the presence of a run that covers position i and the presence of a run that covers position $i+1$ are two events that are most definitely not independent! Linearity of expectation, remember, does not depend on the random variables being independent. Now, for every i, the eight events inside the probability expression are independent of each other, so we can write:

$$p(X_{i-1} \neq 0, X_i = 0, X_{i+1} = 0, X_{i+2} = 0, X_{i+3} = 0, X_{i+4} = 0, X_{i+5} = 0, X_{i+6} \neq 0) \tag{13}$$

$$= p(X_{i-1} \neq 0) \cdot \left(\prod_{j=0}^{5} \cdot p(X_{i+j} = 0) \right) \cdot p(X_{i+6} \neq 0) \tag{14}$$

$$= p(X_{i-1} \neq 0) \cdot \frac{1}{64} \cdot p(X_{i+6} \neq 0) \tag{15}$$

Note that line 15 holds because we know the coin is fair and $(\frac{1}{2})^6 = \frac{1}{64}$. There are three separate cases we need to handle to get the sum in expression 12.

1. When $i = 1$, the first factor $p(X_0 \neq 0) = 1$ (because it's before the beginning of the sequence), and the last factor, $p(X_6 \neq 0) = \frac{1}{2}$. So expression 15 is $\frac{1}{128}$.
2. When $i = 95$, the first factor $p(X_{94} \neq 0) = \frac{1}{2}$, and the last factor, $p(X_{101} \neq 0) = 1$ (because it's past the end of the sequence). So expression 15 is $\frac{1}{128}$.
3. For all other values of i, from 2 to 94, both the first and last factors are $\frac{1}{2}$, so expression 15 is $\frac{1}{256}$. There are 93 cases like this.

So the final answer will be:

$$\frac{1}{128} + 93 \cdot \frac{1}{256} + \frac{1}{128} = \frac{97}{256} \approx 0.3789 \tag{16}$$

Note that, if we had been less careful about the beginning and the end of the sequence, and treated the first and last summands like the middle ones, we'd have come up just a little short (around 0.37). If we'd done that and also ignored the requirement that a proper run of six zeroes must not be preceded or succeeded immediately by another zero, then we'd have come up with a much larger number ($\frac{95}{64} \approx 1.48$).

Bias Detection Let's introduce two random variables. Let X range over $\{\text{biased}, \text{fair}\}$ indicate which coin you were handed. Let $Y_1, \ldots, Y_5$ correspond to the five coin tosses. For the first question, we are interested in $p(X = \text{biased} \mid Y_1 = 0, Y_2 = 0, Y_3 = 1, Y_4 = 0, Y_5 = 1)$. By the definition of conditional probability (and letting y_i denote the observed value of Y_i),

$$p(X = \text{biased} \mid Y_1 = 0, Y_2 = 0, Y_3 = 1, Y_4 = 0, Y_5 = 1) \tag{17}$$

$$= \frac{p(X = \text{biased}, Y_1 = 0, Y_2 = 0, Y_3 = 1, Y_4 = 0, Y_5 = 1)}{\sum_{x \in \{\text{biased,fair}\}} p(X = x, Y_1 = 0, Y_2 = 0, Y_3 = 1, Y_4 = 0, Y_5 = 1)} \tag{18}$$

$$= \frac{p(X = \text{biased}) \cdot \prod_{i-1}^{5} p(Y_i = y_i \mid X = \text{biased})}{p(X = \text{biased}) \cdot \prod_{i-1}^{5} p(Y_i = y_i \mid X = \text{biased}) + p(X = \text{fair}) \cdot \prod_{i-1}^{5} p(Y_i = y_i \mid X = \text{fair})} \tag{19}$$

$$= \frac{\frac{1}{2} \cdot \left(\frac{3}{4}\right)^3 \cdot \left(\frac{1}{4}\right)^2}{\frac{1}{2} \cdot \left(\frac{3}{4}\right)^3 \cdot \left(\frac{1}{4}\right)^2 + \frac{1}{2} \cdot \left(\frac{1}{2}\right)^5} = \frac{\frac{27}{2048}}{\frac{27}{2048} + \frac{1}{64}} \approx 0.4576 \tag{20}$$

The sensible answer would be to guess that you were handed the biased coin. This may come as a surprise, because according to the above calculation, the posterior probability is higher for the fair coin (i.e., $1 - 0.4576 > 0.4576$). Your "maximum *a posteriori*" guess should be the fair coin. But the biased coin is worth more. Note that, if you guess "fair," your expected winnings are

$$0.4576 \times \$0 + (1 - 0.4576) \times \$1000 = \$542.37 \tag{21}$$

because there's still a 45.76% chance you're wrong and will get nothing. Meanwhile, if you guess "biased," then your expected winnings are

$$0.4576 \times \$1200 + (1 - 0.4576) \times \$0 = \$549.15 \tag{22}$$

By a slim margin, you'll do better in expectation by guessing "biased."

Random Diagonal

- Because $A_{1,1}$ is a continuous random variable, the probability that it takes value exactly 0 is zero. (If we'd asked for $p(-\epsilon < A_{1,1} < \epsilon)$, then your answer would have been ϵ.)
- Each $A_{i,i}$ is independent of the others, so

$$p(A_{n,n} > 0.5 \mid A_{n-1,n-1} \geq 0) = p(A_{n,n} > 0.5) \tag{23}$$

$$= \int_{0.5}^{1} \frac{1}{1 - (-1)} dx \tag{24}$$

$$= \frac{1}{2} - \frac{1}{4} \tag{25}$$

$$= \frac{1}{4} \tag{26}$$

- Recall that to find $\text{rank}(\mathbf{A})$, we must determine the number of linearly independent columns it has. The only way column i will become linearly dependent on other columns is if $A_{i,i}$

takes value zero.

$$p(\text{rank}(\mathbf{A}) < n) = p(\text{at least one } A_{i,i} \text{ takes value zero}) \tag{27}$$
$$= 1 - p(\text{all } A_{i,i} \text{ are nonzero}) \tag{28}$$
$$= 1 - \prod_{i=1}^{n} (1 - p(A_{i,i} = 0)) \tag{29}$$
$$= 0 \tag{30}$$

(Refer to the first part of this problem.)

- Recall that $\mathbf{A} \succeq 0$ when all eigenvalues of $\mathbf{A}$ are nonnegative.

  We can find the eigenvalues λ_i of $\mathbf{A}$ by solving the equation[2] $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. This leads to:

$$\begin{vmatrix} a_{1,1} - \lambda_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{n,n} - \lambda_n \end{vmatrix} = 0 \tag{31}$$
$$\prod_{i=1}^{n} (a_{i,i} - \lambda_i) = 0 \tag{32}$$
$$\forall i, \lambda_i = a_{i,i} \tag{33}$$

  Therefore,

$$p(\mathbf{A} \succeq 0) = p(A_{1,1} \geq 0, \dots, A_{n,n} \geq 0) \tag{34}$$
$$= \prod_{i=1}^{n} p(A_{i,i} \geq 0) \tag{35}$$
$$= \left(\frac{1}{2}\right)^n \tag{36}$$

Matrix Operations We first find the eigenvalues of $\mathbf{B}$ by solving:

$$\begin{vmatrix} 1 - \lambda & -1 & 0 \\ -1 & 2 - \lambda & -1 \\ 0 & -1 & 1 - \lambda \end{vmatrix} = 0 \tag{37}$$
$$= -3\lambda + 4\lambda^2 - \lambda^3 \tag{38}$$
$$= \lambda(-3 + 4\lambda - \lambda^2) \tag{39}$$
$$= -\lambda(\lambda - 1)(\lambda - 3) \tag{40}$$

So the eigenvalues of $\mathbf{B}$ are 0, 1, and 3. By the invertible matrix theorem, this lets us conclude that $\mathbf{B}$ is not invertible; it has zero as an eigenvalue.

[2] $\det(\mathbf{A}) = \prod_{i=1}^{n} a_{i,i}$ holds for any diagonal matrix $\mathbf{A}$, which can be shown by cofactor expansion.

Next, we find those three eigenvalues' respective eigenvectors by solving $\mathbf{B} - \lambda\mathbf{x} = \mathbf{0}$. To do this, you can use Gaussian elimination to convert the left side matrix to row echelon form, and then back-substitute. You should get, for eigenvalues 0, 1, and 3 respectively:

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad \mathbf{v}_3 = \begin{bmatrix} 1 \\ -2 \\ 1 \end{bmatrix} \tag{41}$$

Note that any of the above can be multiplied by a nonzero scalar.

Recall that an $n \times n$ matrix with n linearly independent eigenvectors can be diagonalized.

$$\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}, \text{where } \mathbf{X} = [\mathbf{v}_1 \mid \mathbf{v}_2 \mid \mathbf{v}_3] \tag{42}$$

Derivatives of Activation Functions First, the derivative of the sigmoid function:

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}(1 + e^{-x})^{-1} \tag{43}$$

$$= -(1 + e^{-x})^{-2} \cdot \frac{d}{dx}(1 + e^{-x}) \tag{44}$$

$$= -(1 + e^{-x})^{-2} \cdot (-e^{-x}) \tag{45}$$

$$= \frac{e^{-x}}{(1 + e^{-x})^2} \tag{46}$$

$$= \frac{e^{-x}}{1 + e^{-x}} \cdot \frac{1}{1 + e^{-x}} \tag{47}$$

Observe that:

$$1 - \frac{1}{1 + e^{-x}} = \frac{1 + e^{-x}}{1 + e^{-x}} - \frac{1}{1 + e^{-x}} \tag{48}$$

$$= \frac{e^{-x}}{1 + e^{-x}} \tag{49}$$

So we have

$$\frac{d}{dx}\sigma(x) = \frac{e^{-x}}{1 + e^{-x}} \cdot \frac{1}{1 + e^{-x}} \tag{50}$$

$$= \left(1 - \frac{1}{1 + e^{-x}}\right) \cdot \sigma(x) \tag{51}$$

$$= (1 - \sigma(x))) \cdot \sigma(x) \tag{52}$$

For the hyperbolic tangent, there's more than one way to do it; we'll show two of them. First

write $\tanh$ in terms of σ, x, and constants.

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{53}$$
$$= \frac{1 - e^{-2x}}{1 + e^{-2x}} \cdot \frac{e^x}{e^x} \tag{54}$$
$$= \frac{1}{1 + e^{-2x}} \frac{1 - e^{-2x}}{1} \tag{55}$$
$$= a \cdot b \tag{56}$$

Observe that:

$$a = \sigma(2x) \tag{57}$$

$$-b = e^{-2x} - 1 \tag{58}$$
$$-b + 2 = e^{-2x} + 1 \tag{59}$$
$$\frac{1}{-b+2} = \frac{1}{1 + e^{-2x}} = \sigma(2x) \tag{60}$$
$$b = 2 - \frac{1}{\sigma(2x)} \tag{61}$$

Thus

$$\tanh(x) = a \cdot b \tag{62}$$
$$= \sigma(2x)\left(2 - \frac{1}{\sigma(2x)}\right) \tag{63}$$
$$= 2\sigma(2x) - 1 \tag{64}$$

We can then use the chain rule to find its derivative, given that $\frac{d}{du}\sigma(u) = \sigma(u)(1 - \sigma(u))u'$:

$$\frac{d}{dx}\tanh(x) = \frac{d}{dx}[2\sigma(2x) - 1] \tag{65}$$
$$= 2\sigma(2x)(1 - \sigma(2x)) \cdot \frac{d}{dx}[2x] \tag{66}$$
$$= 4\sigma(2x)(1 - \sigma(2x)) \tag{67}$$
$$= 4\sigma(2x) - 4\sigma^2(2x) \tag{68}$$

Another way to do it is to derive $\tanh$ first:

$$\frac{d}{dx}\tanh(x) = \frac{d}{dx}\frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{69}$$

$$= \frac{(e^x + e^{-x})\frac{d}{dx}(e^x - e^{-x}) - (e^x - e^{-x})\frac{d}{dx}(e^x + e^{-x})}{(e^x + e^{-x})^2} \tag{70}$$

$$= \frac{(e^x + e^{-x})(e^x + e^{-x}) - (e^x - e^{-x})(e^x - e^{-x})}{(e^x + e^{-x})^2} \tag{71}$$

$$= \frac{(e^x + e^{-x})^2 - (e^x - e^{-x})^2}{(e^x + e^{-x})^2} \tag{72}$$

$$= \frac{(e^x + e^{-x})^2}{(e^x + e^{-x})^2} - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} \tag{73}$$

$$= 1 - \frac{(e^x - e^{-x})^2}{(e^x + e^{-x})^2} \tag{74}$$

$$= 1 - \tanh^2(x) \tag{75}$$

$$= (1 + \tanh(x))(1 - \tanh(x)) \tag{76}$$

$$= (1 + 2\sigma(2x) - 1)(1 - 2\sigma(2x) + 1) \tag{77}$$

$$= (2\sigma(2x)) \cdot (2 - 2\sigma(2x)) \tag{78}$$

$$= 4\sigma(2x) - 4\sigma^2(2x) \tag{79}$$

Tile Collection Our objective is to travel from $(1, 1)$ to (n, n) while maximizing the total reward. Let $m_{i,j}$ denote the maximum reward you can obtain at (i, j); the goal of the problem is to find $m_{n,n}$. Since we can only move down or right, there are only two tiles that can directly reach (n, n): $(n-1, n)$ and $(n, n-1)$. This means that:

$$m_{n,n} = r_{n,n} + \max(m_{n-1,n}, m_{n,n-1}) \tag{80}$$

The fact that solving the problem at (n, n) reduces to solving two similar but slightly smaller problems is known as "optimal substructure." Next notice that the equation generalizes to all positions on the grid except at the top boundary ($i = 1$), right boundary ($j = 1$), and the top left corner ($i = j = 1$).

$$m_{i,j} = r_{i,j} + \begin{cases} \max(m_{i-1,j}, m_{i,j-1}) & \text{if } i > 1 \wedge j > 1 \\ m_{i,j-1} & \text{if } i = 1 \wedge j > 1 \\ m_{i-1,j} & \text{if } i > 1 \wedge j = 1 \\ 0 & \text{if } i = j = 1 \end{cases} \tag{81}$$

In order to solve the problem, one must start at the base case ($i = j = 1$) and proceed outward to calculate all of $m_{i,j}$. Every time a "max" is calculated, one must record the preceding position that gave the max (known sometimes as the "argmax"). For completeness, here are the argmax

calculations, to be done alongside the $m_{i,j}$ calculations above:

$$a_{i,j} = \begin{cases} (i-1,j) & \text{if } i > 1 \wedge j > 1 \wedge m_{i-1,j} \geq m_{i,j-1} \\ (i,j-1) & \text{if } i > 1 \wedge j > 1 \wedge m_{i-1,j} < m_{i,j-1} \\ (i,j-1) & \text{if } i = 1 \wedge j > 1 \\ (i-1,j) & \text{if } i > 1 \wedge j = 1 \\ \emptyset & \text{if } i = j = 1 \end{cases} \tag{82}$$

By following back the trail of argmaxes from (n, n) to $(1, 1)$, one recovers the best path in reverse. The reward, again, is $m_{n,n}$.

Space complexity is $O(n^2)$, because we need to store a constant amount of information for each tile on the grid ($m_{i,j}$ and $a_{i,j}$). Runtime is also $O(n^2)$; we perform a constant amount of work for each tile.

Lamp and Box Most people will understand *it* to refer to the lamp in the first example and the box in the second example. Here's another example (who does *they* refer to?):

1. The city councilmen refused the demonstrators a permit because they feared violence.

2. The city councilmen refused the demonstrators a permit because they advocated violence.

These judgments are generally considered to rely on commonsense knowledge; we need to know that, in general, for X to fit in Y, X should be smaller than Y.

To learn more about the history of these kinds of questions, see `https://en.wikipedia.org/wiki/Winograd_schema_challenge`. A more recent version of the challenge is discussed in this paper, which was published at the AAAI 2020 conference: `https://arxiv.org/pdf/1907.10641.pdf`.