# Math Crash Course

A brief overview of probability, statistical inference, vector algebra, and optimization

Kyle Fridberg

July 15, 2021

# Outline


1. Probability
2. Statistical Inference
3. Vectors
4. Optimization

# Set Notation Review

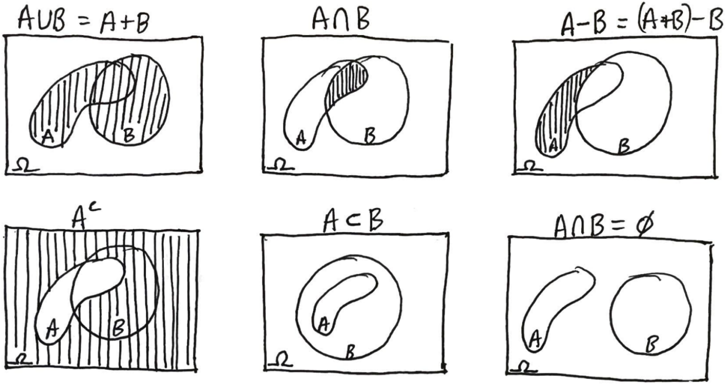
**Why care about sets?**

- Foundations of mathematics are built upon set theory.
  - Zermelo–Fraenkel axioms (ZF/ZFC).
- Connection between math and philosophy.
- Probability theory = chance of event occurring within set of events that are possible.

# Set Notation Review

**Why care about sets?**

- Foundations of mathematics are built upon set theory.
  - Zermelo–Fraenkel axioms (ZF/ZFC).
- Connection between math and philosophy.
- Probability theory = chance of event occurring within set of events that are possible.



# Probability Basics

**Sample space:** Set of all possible outcomes of an experiment ($\Omega$).

**Ex 1:** Tossing coin twice, $\Omega = \{HH, HT, TH, TT\}$.

**P(X):** Probability of event X happening.

**Ex 2:** Tossing coin twice, $P(HH) = P(H) \cdot P(H) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. What does this mean (philosophically)?

1. Frequentist: P(HH) expresses that if we flip a coin twice many many times, the number of times we observe HH will approach $\frac{1}{4}$ of the total.
2. Bayesian: P(HH) expresses our confidence that we will roll HH in a given trial.

*Everything lives in a sample space, even if we don't talk about it!*

# Counting

**Permutation:** Number of ways to choose elements from a set to create ordered lists.

**Ex 3:** How many ways to arrange the letters of the word OCTOPUS such that the two Os stay together and the S stays at the end?

# Counting

**Permutation:** Number of ways to choose elements from a set to create ordered lists.

**Ex 3:** How many ways to arrange the letters of the word OCTOPUS such that the two Os stay together and the S stays at the end?

$$\text{OCTOPUS} \rightarrow \{\underset{1}{OO}, \underset{2}{C}, \underset{3}{T}, \underset{4}{P}, \underset{5}{U}, \cancel{S}\} \rightarrow \underset{\text{5 choices}}{\_} \ \underset{\text{4 choices}}{\_} \ \underset{3}{\_} \ \underset{2}{\_} \ \underset{1}{\_} \ \underline{S}$$

$$= 5! = \boxed{120.}$$

# Counting cont.

**Combination:** Number of ways to choose elements from a set to create subsets. Order doesn't matter.

**Ex 4:** How many subsets of $\{1, 2, 3, 4, 5\}$ have a sum greater than $\pi$?

# Counting cont.

**Combination:** Number of ways to choose elements from a set to create subsets. Order doesn't matter.

**Ex 4:** How many subsets of $\{1, 2, 3, 4, 5\}$ have a sum greater than $\pi$?

Hint: the total number of subsets (including the empty set!) of a set with $n$ elements is $2^n$. The set of all possible subsets of a set $S$ is called the power set of $S$.

# Counting cont.

**Combination:** Number of ways to choose elements from a set to create subsets. Order doesn't matter.

**Ex 4:** How many subsets of $\{1, 2, 3, 4, 5\}$ have a sum greater than $\pi$?

Hint: the total number of subsets (including the empty set!) of a set with $n$ elements is $2^n$. The set of all possible subsets of a set $S$ is called the power set of $S$.

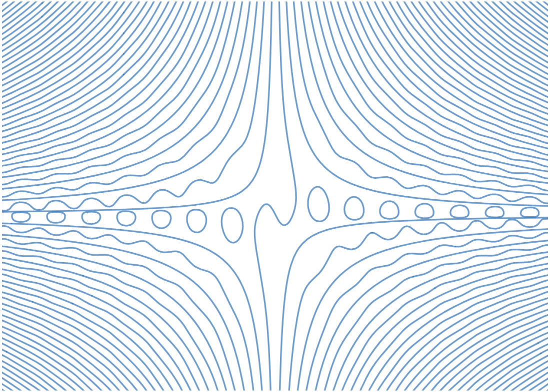
$S = \{1, 2, 3, 4, 5\}$. Count all subsets of $S$ such that the sum of the elements in the subset is greater than $\pi$.

We instead count subsets with sum $< \pi$, then subtract from total.

Subsets with sum $< \pi$: $\{1\}, \{2\}, \{3\}, \{1,2\}, \emptyset$ = 5 subsets.

Total # subsets $= 2^5 = 32$. Answer: $32 - 5 = \boxed{27}$

# Discussion question



Consider the power set of $\mathbb{R}^2$ (the $x$-$y$ plane). What types of things does this set contain?

# Independence vs Dependence

**Def:** Events A and B are independent if and only if $P(A, B) = P(A) \cdot P(B)$.

**Ex 5:** A royal flush in Poker is when you get $\{10, J, Q, K, A\}$ all of the same suit. What is the probability of someone dealing you a royal flush from a standard deck of 52 cards?

# Independence vs Dependence

**Def:** Events A and B are independent if and only if
$P(A, B) = P(A) \cdot P(B)$.

**Ex 5:** A royal flush in Poker is when you get $\{10, J, Q, K, A\}$ all of the same suit. What is the probability of someone dealing you a royal flush from a standard deck of 52 cards?

$\underline{J\heartsuit}$ $\underline{10\heartsuit}$ $\underline{K\heartsuit}$ $\underline{Q\heartsuit}$ $\underline{A\heartsuit}$ $= \frac{20}{52} \cdot \frac{4}{51} \cdot \frac{3}{50} \cdot \frac{2}{49} \cdot \frac{1}{48}$

10, J, Q, K, A of 4 suits $= 5 \cdot 4 = 20$ possibilities out of 52

10, Q, K, A of hearts = 4 poss. out of 51

Q, K, A = 3 poss. out of 50

Q, A = 2 poss out of 49

A = 1 poss out of 48

$= \frac{1}{649{,}740}$

probabilities dependent on # of cards dealt.

# Independence vs Dependence

**Def:** Events A and B are independent if and only if $P(A, B) = P(A) \cdot P(B)$.

**Ex 5:** A royal flush in Poker is when you get $\{10, J, Q, K, A\}$ all of the same suit. What is the probability of someone dealing you a royal flush from a standard deck of 52 cards?

J♥ 10♥ K♥ Q♥ A♥

$$= \frac{20}{52} \cdot \frac{4}{51} \cdot \frac{3}{50} \cdot \frac{2}{49} \cdot \frac{1}{48} = \frac{1}{649{,}740}$$

10, J, Q, K, A of 4 suits = 5·4 = 20 possibilities out of 52

10, Q, K, A of hearts = 4 poss. out of 51

Q, K, A = 3 poss. out of 50

Q, A = 2 poss out of 49

A = 1 poss out of 48

probabilities dependent on # of cards dealt.

From an infinite number of decks of cards?

# Independence vs Dependence

**Def:** Events A and B are independent if and only if $P(A, B) = P(A) \cdot P(B)$.

**Ex 5:** A royal flush in Poker is when you get $\{10, J, Q, K, A\}$ all of the same suit. What is the probability of someone dealing you a royal flush from a standard deck of 52 cards?

$\underline{J\heartsuit}\ \ \underline{10\heartsuit}\ \ \underline{K\heartsuit}\ \ \underline{Q\heartsuit}\ \ \underline{A\heartsuit} = \frac{20}{52} \cdot \frac{4}{51} \cdot \frac{3}{50} \cdot \frac{2}{49} \cdot \frac{1}{48}$

10, J, Q, K, A of 4 suits $= 5 \cdot 4 = 20$ possibilities out of 52

10, Q, K, A of hearts = 4 poss. out of 51

Q, K, A = 3 poss. out of 50

Q, A = 2 poss out of 49

A = 1 poss out of 48

$= \frac{1}{649{,}740}$

probabilities dependent on # of cards dealt.

From an infinite number of decks of cards?

Infinite decks of cards. Good trick: instead consider $n$ decks of cards, get algebraic expression, and ~~then~~ take limit $n \to \infty$.

$$\lim_{n\to\infty}\left(\frac{20n}{52n} \cdot \frac{4n}{52n-1} \cdot \frac{3n}{52n-2} \cdot \frac{2n}{52n-3} \cdot \frac{n}{52n-4}\right) = \frac{20}{52} \cdot \frac{4}{52} \cdot \frac{3}{52} \cdot \frac{2}{52} \cdot \frac{1}{52} = \frac{15}{11{,}881{,}376} < \frac{1}{649{,}740}$$

probabilities independent of # of cards dealt.

# Conditional Probability

**Def:** Given two events A and B, $P(A|B) = \frac{P(A,B)}{P(B)}$.

We divide by the probability of B occurring because we already know/assume it happened (or is going to happen).

**Prosecutor's fallacy:** In general, $P(A|B) \neq P(B|A)$.

**Ex 6:**
Probability of nightmare | dreaming $< 1$
Probability of dreaming | nightmare $= 1$

# Two of the Most Important Theorems in Probability Theory

**Law of Total Probability (LOTP):**

$$P(A) = \sum_n P(A|B_n)P(B_n)$$

**Bayes Theorem:**

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

# Do YOU have COVID-19?

**Ex 7:** Oh no, you've taken a COVID-19 test and you tested positive! What information/data would be needed to determine the probability that you actually have COVID?

# Do YOU have COVID-19?

**Ex 7:** Oh no, you've taken a COVID-19 test and you tested positive! What information/data would be needed to determine the probability that you actually have COVID?

*Given:*

- $P(D) \approx 0.001$ for 10-17 year olds in Boulder County (as of last Monday)
- $P(+|D) \approx 0.8$ for rapid COVID-19 antigen tests
- $P(+|D^c) \approx 0.005$ false positive rate (assumed)
- Notation: $D =$ have COVID, $D^c =$ don't have COVID, $+ =$ test positive, $- =$ test negative

You will need LOTP: $P(A) = \sum_n P(A|B_n)P(B_n)$

and Bayes Theorem: $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$.

# Do YOU have COVID-19?

*Hint:* Using LOTP, $P(+) = P(+|D)P(D) + P(+|D^c)P(D^c)$.

# Do YOU have COVID-19?

*Hint:* Using LOTP, $P(+) = P(+|D)P(D) + P(+|D^c)P(D^c)$.

Want to calculate $P(D|+)$ using information: $\begin{cases} P(D) = 0.001 \\ P(+|D) = 0.8 \\ P(+|D^c) = 0.005 \end{cases}$

Using Bayes Theorem:

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+)} = \frac{0.8 \cdot 0.001}{P(+)}$$

But what is $P(+)$?

We can get it using LOTP: 
$$P(+) = P(+|D)\,P(D) + P(+|D^c)\,P(D^c)$$
$$= 0.8 \cdot 0.001 + 0.005 \cdot (1 - 0.001)$$
$$\approx 0.00580$$

Then, finally: 
$$P(D|+) = \frac{0.8 \cdot 0.001}{0.00580} \approx 0.138$$

So you only have a 13.8 % chance of having covid!

# Random Variables

**Def:** A random variable is a variable whose value is unknown until after the experiment.

**Ex 8:**
Boring experiment: $X = \text{bool}(\text{name} == \text{“Kyle”})$. Input name...
$P(X = 1) = 0.0001$.
But after experiment, $X = 1$ because my name is Kyle.

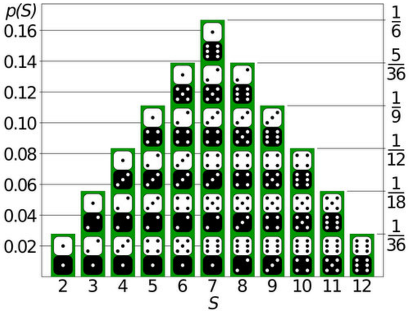
We say that a random variable "crystallizes" (takes on a value) after performing the experiment.

# What is a Probability Distribution?

Q: You will likely see the words PMF, PDF, CDF thrown around... but which one represents the probability distribution?

A: All of them, but in different contexts.

**Ex 9:** Let S be the outcome (sum) of rolling 2 dice in Settlers of Catan. We say $S \sim P(S = y) = p_S(y)$, which is represented visually by the following function. $P(S = 7) = \frac{1}{6}$ (robber), but $P(S = 20) = 0$.

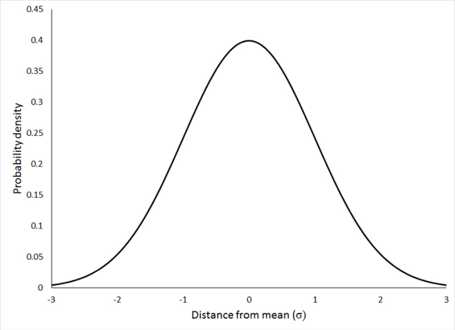


# The Most Famous Probability Distribution

Gaussian (a.k.a. Normal) distribution:

Basic form of Probability Density Function (PDF): $f(x) = e^{-x^2}$.

With some added parameters to be able to modify the mean, variance, and ensure it is a valid probability distribution.

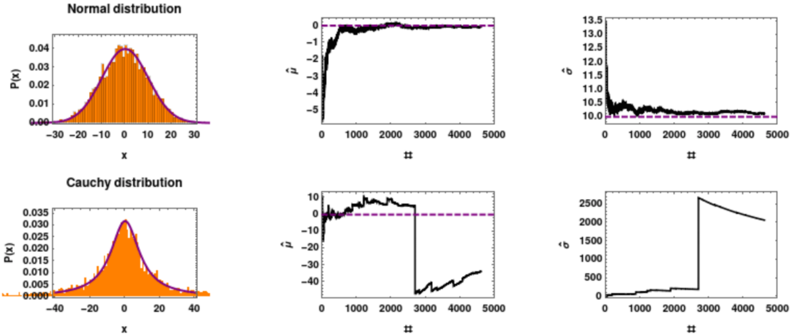


# A Remarkable Probability Distribution

The Cauchy (a.k.a. Lorentz) distribution has a PDF of the following form: $f(x) = \frac{1}{1+x^2}$.
What makes it remarkable is that both its mean and variance are undefined, meaning that no matter how many times we sample from the distribution, the average of the points sampled never converges.

# A Remarkable Probability Distribution

The Cauchy (a.k.a. Lorentz) distribution has a PDF of the following form: $f(x) = \frac{1}{1+x^2}$.
What makes it remarkable is that both its mean and variance are undefined, meaning that no matter how many times we sample from the distribution, the average of the points sampled never converges.



# Statistical Inference (Bird's-Eye View)

**Def:** Statistical inference (a.k.a learning) is using data to infer the distribution that generated the data.

The question we aim to answer is the following:
Given a sample $X_1, X_2, ..., X_n \sim F$, how do we infer F?

# Statistical Inference vs ML

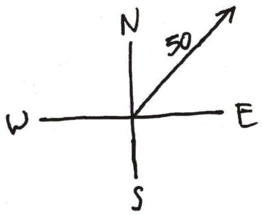
| **Statistical Inference** | **Machine Learning** |
|---|---|
| Infer facts about population from sample | Find generalizable predictive patterns (in large and wide datasets) |
| Includes assumptions about how data is generated | No such assumptions |
| Probabilistic model, p-values, confidence interval | Emphasis on predictive accuracy through cross-validation |
| Excellent for solving theoretical problems | Excellent for solving engineering problems |

*This table is based on a table from the video "Differences Between Statistics & Machine Learning." For more details, watch the video!

# Vector Basics

A **vector** is an object with 2 properties: magnitude and direction.

"50 miles northeast" represents a vector on the map:




Examples of vectors?

# Vector Properties

Using the vectors $\vec{v} = (v_1, v_2, ..., v_n)$ and $\vec{w} = (w_1, w_2, ..., w_n)$

**Magnitude:** $|\vec{v}| = \sqrt{v_1^2 + v_2^2 + ... + v_n^2}$.

**Addition:** $\vec{v} + \vec{w} = (v_1 + w_1, v_2 + w_2, ..., v_n + w_n)$.

**Scalar multiplication:** $c \cdot \vec{v} = (cv_1, cv_2, ..., cv_n)$.

**Dot product:** $\vec{v} \cdot \vec{w} = v_1 w_1 + v_2 w_2 + ... + v_n w_n = |\vec{v}| \cdot |\vec{w}| \cos(\theta)$

where $\theta$ is the angle between $\vec{v}$ and $\vec{w}$.

**Vector between two points:** $P = (a_1, a_2, ...a_n)$ and $Q = (b_1, b_2, ..., b_n)$,

then $\vec{v}_{PQ} = Q - P = (b_1 - a_1, b_2 - a_2, ..., b_n - a_n)$.

# Spacecraft Launch

**Ex 1:** We desire to launch a spacecraft at an angle of 80 degrees to the horizon. Given that the earth is FLAT (obviously) and the atmosphere is 300 miles thick, how many miles must the spacecraft travel until it escapes the atmosphere?

# Spacecraft Launch

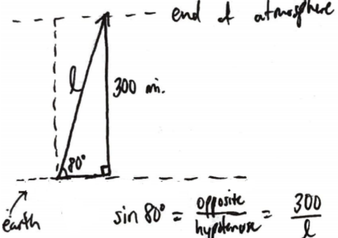
**Ex 1:** We desire to launch a spacecraft at an angle of 80 degrees to the horizon. Given that the earth is FLAT (obviously) and the atmosphere is 300 miles thick, how many miles must the spacecraft travel until it escapes the atmosphere?

Hint: Use SOH-CAH-TOA.

# Spacecraft Launch

# Spacecraft Launch

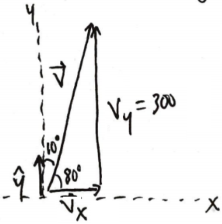
Method 1: Using trigonometry




$$\sin 80^\circ = \frac{\text{opposite}}{\text{hypotenuse}} = \frac{300}{\ell}$$

$$\Rightarrow \boxed{\ell = \frac{300}{\sin 80^\circ} \approx 304.63 \text{ mi.}}$$

Method 2: Using vector dot product.




$$\vec{v} = (v_x, \underset{v_y}{300}).$$

Want to find $|\vec{v}|$

$\hat{y}$ is unit vector in y direction: $\hat{y} = (0,1)$

Angle btwn. $\vec{v}$ and $\hat{y}$ is 10°.

$$\vec{v} \cdot \hat{y} = |\vec{v}| \cdot |\hat{y}| \cos 10^\circ$$

$$(v_x, 300) \cdot (0,1) = |\vec{v}| \cdot 1 \cdot \cos 10^\circ$$

$$300 = |\vec{v}| \cdot \cos 10^\circ$$

$$\boxed{|\vec{v}| = \frac{300}{\cos 10^\circ} \approx 304.63 \text{ mi}}$$

# Spacecraft Launch (continued)

**Challenge problem:** We desire to launch a spacecraft at an angle of 80 degrees to the horizon. Given that the earth is ROUND, the atmosphere is 300 miles thick, and the earth is 8000 miles in diameter, how many miles must the spacecraft travel until it escapes the atmosphere?

# Spacecraft Launch (continued)

**Challenge problem:** We desire to launch a spacecraft at an angle of 80 degrees to the horizon. Given that the earth is ROUND, the atmosphere is 300 miles thick, and the earth is 8000 miles in diameter, how many miles must the spacecraft travel until it escapes the atmosphere?

Intuitively, will the answer be smaller or larger than in the previous example?

# Spacecraft Launch (continued)

**Challenge problem:** We desire to launch a spacecraft at an angle of 80 degrees to the horizon. Given that the earth is ROUND, the atmosphere is 300 miles thick, and the earth is 8000 miles in diameter, how many miles must the spacecraft travel until it escapes the atmosphere?

Intuitively, will the answer be smaller or larger than in the previous example?

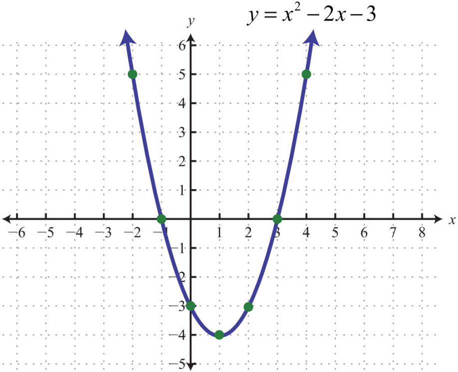
300 miles $<$ answer $<$ 304.63 miles

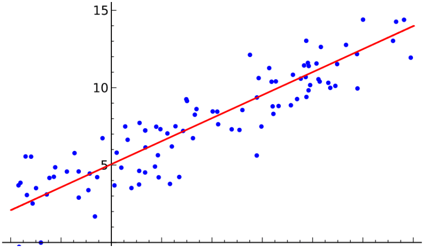
# What is optimization?

**Def:** "Mathematical optimization is the selection of a best element, with regard to some criterion, from some set of available alternatives."

# What is optimization?

**Def:** "Mathematical optimization is the selection of a best element, with regard to some criterion, from some set of available alternatives."



# Real-World Examples

# Real-World Examples

- Google maps (find optimal path between two points on the globe).

# Real-World Examples

- Google maps (find optimal path between two points on the globe).
- Administering medications (treat disease without harming other bodily functions).

# Real-World Examples

- Google maps (find optimal path between two points on the globe).
- Administering medications (treat disease without harming other bodily functions).
- Selling a product (low enough price that many people will buy it, but high enough price that make profit).

# Real-World Examples

- Google maps (find optimal path between two points on the globe).
- Administering medications (treat disease without harming other bodily functions).
- Selling a product (low enough price that many people will buy it, but high enough price that make profit).
- Others?

# Optimization Terms

**Calculus:** Powerful technique for analyzing the rates of change of mathematical functions.

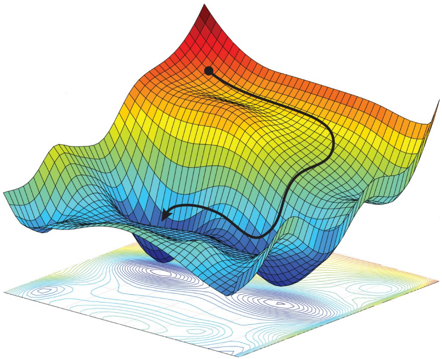
**Gradient descent:** Use gradient (slope) of function to approach local minimum.

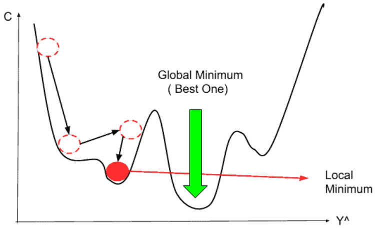
**Stochastic gradient descent (SGD):** Approximate gradient of noisy data set to approach local (or global) minimum. Very important optimization method in ML.

**Loss function / Cost function:** Function that quantifies the difference between the model and the data.

*The best model is the one that minimizes the loss function!*

# Gradient Descent Visualized

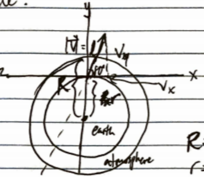




# Solution to Challenge Spacecraft Launch Problem

Just do with circle:

$x^2+y^2=R^2$




① $x^2+(y+4000)^2=4300^2$

② $\vec{V}=(l\cos 80°, l\sin 80°)$

and $|\vec{V}|=l$

$R=4000+300$

$r=4000$

Want to find length $l$ such that $\vec{V}$ intersects with outer circle.

Plugging ② into ①:

$$(l\cos 80°)^2+(l\sin 80°+4000)^2=4300^2$$

$$l^2\cos^2 80°+l^2\sin^2 80°+8000\,l\sin 80°+4000^2=4300^2$$

$$l^2(\underbrace{\cos^2 80°+\sin^2 80°}_{1})+8000\,l\sin 80°+4000^2-4300^2=0.$$

Use quadratic formula to get $l$:

$$l=\frac{-8000\sin 80°\pm\sqrt{(8000\sin 80°)^2-4\cdot 1(4000^2-4300^2)}}{2\cdot 1}$$

$$\approx \boxed{304.298},\ -8182.760$$

Note: If we launched the spacecraft through the earth in the opposite direction, this is how many miles for it to cross atmosphere on other side.