Solving Problems by Rolling Down the Hill

Michael J. Neely
University of Southern California
http://www-rcf.usc.edu/∼mjneely

Abstract

These notes discuss techniques for solving problems that appear in theory courses within mathematics, computational science, and engineering. The idea is that, for any theoretical subject, there are a class of problems that involve a step-by-step application of basic definitions. Such problems can be solved by appropriately "rolling down the hill." These notes provide three examples of rolling down the hill from the subjects of probability, linear algebra, and optimization. They also discuss common mistakes. An exercise that emphasizes the distinction between "there exists" and "for all" is provided at the end.

I. Introduction

There are a class of problems that I like to call *roll-down-the-hill* problems. These problems appear in any theoretical course in mathematics, computational science, or engineering. The problems involve proving or verifying a certain claim. At first, it may not be obvious why the claim is true. However, roll-down-the-hill problems can be solved simply by stating what we are given, then stating what we want to show, and then applying basic definitions step-by-step until we have shown the desired result. Once the problem is set up properly, there is often only one possible step to do next, and then one possible step to do after that, and so on, until we are done. This is similar to setting a ball at the top of a steep hill: If positioned properly, there is only one way to roll down. If we think of the basic definitions as *gravity*, then it becomes impossible *not* to solve the problem once it is set up correctly, unless we want to defy gravity. Below I give examples of this from three different subjects: probability, linear algebra, and optimization. All of these examples are simply examples of mathematical analysis.

II. Examples of Rolling Down the Hill

A. Probability

In a probability class, a common problem is to compute the distribution function of one random variable in terms of the distribution function of another.

Definition 1: The *cumulative distribution function* (CDF) for a random variable X, written $F_X(x)$, is defined:

$$F_X(x) = Pr[X \leq x] \ \forall x \in \mathbb{R}$$

Problem 1: Let X be a random variable with a given CDF $F_X(x)$. Let $Y = 2X + 3$. Find the CDF $F_Y(y)$.

Solution:

- *Set the ball on top of the hill:* We write: "Let $Y = 2X + 3$, where X has CDF $F_X(x)$. We want to find the CDF $F_Y(y)$. By definition, we have $F_Y(y) = Pr[Y \leq y]$."
- *Roll down the hill:* We have:

$$\begin{aligned} F_Y(y) &= Pr[Y \leq y] \\ &= Pr[2X + 3 \leq y] \\ &= Pr[X \leq (y-3)/2] \\ &= F_X((y-3)/2) \end{aligned}$$

We see that there was basically only one way to roll down the hill: Just go step-by-step, applying the definition of CDF and applying the given equation $Y = 2X + 3$. If we choose to ignore either the definition of CDF or the given equation, we *cannot* solve the problem. However, it is almost impossible *not* to solve the problem if we allow ourselves to use these things.

B. Linear Algebra

Linear independence is one of the most important concepts in linear algebra. Every linear algebra course includes homework problems where we are given a collection of vectors that satisfy some property, and we must prove these vectors are linearly independent.

Definition 2: A collection of vectors $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are *linearly independent* if the equation $\sum_{i=1}^n \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$ can only be solved when the scalars α_i satisfy $\alpha_i = 0$ for all $i \in \{1, \ldots, n\}$.

Problem 2: Suppose $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ are a collection of linearly independent vectors. Let A be a matrix, and suppose that for each $i \in \{1, \ldots, n\}$ there is a vector $\boldsymbol{x}_i$ such that $A\boldsymbol{x}_i = \boldsymbol{y}_i$. Prove that $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ are linearly independent.

Solution:

- *Set the ball on top of the hill:* Write "Suppose that $\sum_{i=1}^n \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$ for some scalars α_i. We want to show that $\alpha_i = 0$ for all i."
- *Roll down the hill:* A correct use of the definition of linearly independent has put us at the top of the hill. We now simply note: If $\sum_{i=1}^n \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$, we can multiply both sides by A to get:

$$A \sum_{i=1}^n \alpha_i \boldsymbol{x}_i = A\boldsymbol{0}$$

and hence:

$$\sum_{i=1}^n \alpha_i A\boldsymbol{x}_i = \boldsymbol{0}$$

and hence, using the definition of $\boldsymbol{y}_i$:

$$\sum_{i=1}^n \alpha_i \boldsymbol{y}_i = \boldsymbol{0}$$

Now, using the fact that $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ are linearly independent, we know the above equation can only be true if $\alpha_i = 0$ for all $i \in \{1, \ldots, n\}$. This proves the result.

Note that once we had the equation $\sum_{i=1}^n \alpha_i \boldsymbol{x}_i = \boldsymbol{0}$, there was nothing left to do except multiply by A. So, in a sense, there was only one possible way to do each step of the proof. Thus, if we get stuck, we just go in the only possible direction that is open. These roll-down-the-hill problems are usually fun, and it is almost impossible to get them wrong if we just correctly apply the definitions.

C. Optimization

An important fact of optimization theory is that the feasibility region of any linear program is convex.

Definition 3: A set $\mathcal{X}$ in $\mathbb{R}^N$ is said to be *convex* if for any two points $\boldsymbol{x}$, $\boldsymbol{y}$ in $\mathcal{X}$ and any value θ such that $0 \leq \theta \leq 1$, we have:

$$\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{X}$$

Problem 3: Let $\mathcal{X}$ be the set of all vectors $\boldsymbol{x} \in \mathbb{R}^N$ that satisfy the entry-wise inequality constraint:

$$A\boldsymbol{x} \leq \boldsymbol{b}$$

where A and $\boldsymbol{b}$ are a given real-valued matrix and vector, respectively, with appropriate sizes so that the multiplication $A\boldsymbol{x}$ makes sense and the entry-wise inequality $A\boldsymbol{x} \leq \boldsymbol{b}$ makes sense. Prove that the set $\mathcal{X}$ is convex.

Solution:

- *Set the ball on top of the hill:* Write "Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two vectors in $\mathcal{X}$, and let θ be a value such that $0 \leq \theta \leq 1$. We want to show that $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{X}$. That is, we want to show that $A(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \boldsymbol{b}$."
- *Roll down the hill:* We have by basic linear algebra:

$$A(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) = \theta A\boldsymbol{x} + (1-\theta)A\boldsymbol{y}$$

However, since $\boldsymbol{x}$ and $\boldsymbol{y}$ are in the set $\mathcal{X}$, we know that $A\boldsymbol{x} \leq \boldsymbol{b}$ and $A\boldsymbol{y} \leq \boldsymbol{b}$. Using these inequalities in the above equation yields:

$$A(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y}) \leq \theta\boldsymbol{b} + (1-\theta)\boldsymbol{b} = \boldsymbol{b}$$

which completes the proof.

III. Common Mistakes

I suspect all professors secretly agree that, in principle, it is impossible *not* to solve such problems that involve rolling down the hill. However, in practice, all professors know that *students often find creative ways to fail*. Below, I discuss four common ways that someone can fail: (i) Failure due to incorrect definitions, (ii) Failure due to confusing what we want to prove with what is given, (iii) Failure due to incorrect handling of given information, (iv) Failure due to rebellious behavior.

A. Failure due to incorrect definitions

Failure is not surprising if incorrect definitions are used. Often students use incorrect definitions because they have a muddled understanding of the correct definition. Other times, students feel that the correct definition is too simple. This is often the case if the student believes the class should be difficult, and thus seeks to complexify definitions in accordance with this belief. As an example, consider the definition of a convex set from the previous section, provided again below for convenience:

Correct Definition of a Convex Set: A set $\mathcal{X}$ in $\mathbb{R}^N$ is said to be *convex* if for any two points $\boldsymbol{x}$, $\boldsymbol{y}$ in $\mathcal{X}$ and any value θ such that $0 \leq \theta \leq 1$, we have $\theta\boldsymbol{x} + (1-\theta)\boldsymbol{y} \in \mathcal{X}$.

For intuition, it is useful to notice that this definition implies that a set is convex if for any two points in the set, the line segment between the points is also in the set. Below I list a collection of *incorrect* definitions of a convex set. These were all actual answers given by students on an exam in my optimization class. Compare their complexity and clarity to that of the true definition.

- Answer 1: *A set $\mathcal{X} \subseteq \mathbb{R}^N$ is a convex set if for all $\boldsymbol{x}$, $\boldsymbol{y}$ in $\mathcal{X}$, there exists a $\lambda \in [0,1]$ such that $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{X}$.*

- Answer 2: *A set $\mathcal{X}$ is a convex set if for all vectors $\boldsymbol{x} \in \mathcal{X}$, there exists some vectors $\boldsymbol{p} \in \mathcal{X}$, $\boldsymbol{q} \in \mathcal{X}$ and a scalar value $0 \leq \theta \leq 1$ such that $\boldsymbol{p}\theta + \boldsymbol{q}(1-\theta) = \boldsymbol{x}$.*

- Answer 3: *$\mathcal{X}$ is a convex set if there are N vectors within the set that satisfy $\sum_{i=1}^N p_i\boldsymbol{x}_i \leq \boldsymbol{b}$, where $\sum_{i=1}^N p_i = 1$ and $\boldsymbol{x}_i$ is a vector in $\mathcal{X}$ for all $i \in \{1,\ldots,N\}$ where $\boldsymbol{b}$ is a vector in $\mathbb{R}^N$.*

- Answer 4: *A convex set $\mathcal{X}$ is defined as a set of all points $\boldsymbol{x}$ such that for a given k points $\boldsymbol{x}_i$, $i \in \{1,\ldots,k\}$ all belonging to $\mathcal{X}$, we can write $\boldsymbol{x} = \sum_{i=1}^k p_i\boldsymbol{x}_i$ where $p_i \geq 0$ for all $i \in \{1,\ldots,k\}$ and $\sum_{i=1}^k p_i = 1$.*

The first two answers are stated precisely, yet the definitions are muddled.[1] It is not difficult to show that the properties stated in these first two answers hold for *all* sets $\mathcal{X}$, not just for convex sets.[2] The third answer is complexified beyond understanding. The most mind boggling is the fourth answer. Rather than defining a property that a given set $\mathcal{X}$ should have, it seems to define a *new* set, also called $\mathcal{X}$, recursively in terms of itself. I sometimes wonder if students think as follows: "I didn't quite understand the correct definition when it was given in class. Therefore, if I write down something that I don't quite understand, it is likely to be correct."

B. Failure due to confusing what we want to prove with what we are given

Consider the linear algebra problem of Section II-B. This problem seeks to prove that a given set of vectors $\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\}$ is linearly independent. It is correctly approached by starting: "Suppose that $\sum_{i=1}^n \alpha_i\boldsymbol{x}_i = \boldsymbol{0}$." However, the most common mistake that students make on this problem is to start backwards by saying: "Suppose

[1]The difference between "there exists" and "for all" is explored in Section V. Answer 1 can be corrected by changing the ending from "there exists a $\lambda \in [0,1]$ such that $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{X}$" to "and for all $\lambda \in [0,1]$, we have $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in \mathcal{X}$."

[2]For example, just take $\lambda = 1$ for answer 1, and $\theta = 1$ for answer 2. The answers 1 and 2 can be made more interesting by removing the $\lambda = 1$ and $\theta = 1$ option. The resulting modified property for answer 1 would be that for any two distinct points in $\mathcal{X}$, there is another distinct point in $\mathcal{X}$ on the line segment between them. The resulting modified property for answer 2 would be that for any point in $\mathcal{X}$, there are two other distinct points in $\mathcal{X}$ for which the first point is in the line segment between the other two. It is not difficult to give examples of non-convex sets that satisfy both of these modified properties.

$\sum_{i=1}^{n} \alpha_i \boldsymbol{y}_i = \mathbf{0}$." This makes it seem like we are trying to prove that the $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ vectors are linearly independent. But we are *given* that these are linearly independent. This backwards approach tries to "prove" that which is already given. This is an exercise in futility that cannot go anywhere meaningful.

C. Failure due to incorrect handling of given information

A mistake closely related to the above is that of either ignoring or incorrectly using a given piece of information. For example, in any probability class, there are a collection of roll-down-the-hill problems where we are given a random variable X with behavior that is contingent on the occurrence of a particular event within a collection of mutually exclusive, collectively exhaustive events. Typically the problem is worded so as to directly provide the student with the conditional probability distribution of X, given each particular event in the collection. A correct way to roll down the hill is then to use the weighted sum of these conditional distributions (weighted by the probability of each corresponding event). However, students often get off track by trying to "derive" the conditional distribution via the (correct) definition of conditional probability $Pr[\mathcal{A}|\mathcal{B}] = Pr[\mathcal{A} \cap \mathcal{B}]/Pr[\mathcal{B}]$. This definition is certainly true, but is often irrelevant to the problem. That is because we are *already given* the conditional distribution we need. We want to go forwards with this information in hand, rather than backwards by trying to prove that which is already given.

D. Failure due to rebellious behavior

When I teach these subjects, I emphasize that students should start their proofs by writing what they are given, and then writing what they want to show. This is an effort to teach them to put the ball at the top of the hill (after which I hope they can roll down). Nevertheless, there are always some students who refuse. Some refuse because they feel it is good enough to "say" or "think about" these initial steps, but would never take the time to write them down. Others refuse by only superficially carrying out my recommendations. For example, they might write something useless such as "we are given the information in the problem above. We want to show that the problem can be solved." Others refuse simply because they want to go their own way.

IV. Clever Tricks

Not all problems are "roll-down-the-hill" problems. Some require more thinking. Some require clever tricks. However, in classes where there are clever tricks, there are usually only a small handful of such tricks, and these can be applied again and again for different permutations of problems. So, it is usually sufficient to just master the "roll-down-the-hill" technique, together with the standard tricks that will surely be demonstrated in class. Sometimes a professor will design an exam problem that involves a part that is "roll-down-the-hill" and a part that uses one of the established clever tricks.

One may want to know the most important clever tricks up a mathematician's sleeve. Here they are:

- Trick 1: Add and subtract the same thing in an equation.
- Trick 2: Multiply and divide the same (non-zero) thing in an equation.

While these seem more like algebra manipulations than tricks, they in fact can be used to solve most real analysis problems. It is not easy to perform these tricks in the correct manner. Mastering them is an art form. One must observe the equation in hand and wish for something better. To act on this wish, one must manipulate the equation, like a sculpture, to make it into that desired object.

A. A Renewal Theory Example

Recall that the law of large numbers proves that the empirical average of a sequence of independent and identically distributed (i.i.d.) random variables $\{X_1, X_2, X_3, \ldots\}$ converges to the expectation $\mathbb{E}[X_1]$ with probability 1. Closely related to sequences of i.i.d. random variables are a class of stochastic processes called *renewal-reward processes*. These are defined over frames $k \in \{1, 2, 3, \ldots\}$, where each frame has a *frame size* T_k and a *reward* R_k. All frame sizes T_k are assumed to be positive. The *time average reward* can then be defined as the limit of the ratio of total rewards to total time:

$$\text{time average reward} = \lim_{K \to \infty} \frac{\sum_{k=1}^{K} R_k}{\sum_{k=1}^{K} T_k}$$

where the above definition assumes the limit exists, and samples the stochastic process only at frame boundaries for simplicity. One typically wants to know when this limit indeed exists, and what its value is. This is elegantly solved by imposing i.i.d. structure on the problem.

Theorem 1: If the vectors (T_k, R_k) are i.i.d. over frames $k \in \{1, 2, 3, \ldots\}$, with finite expectations $\mathbb{E}[T_1]$ and $\mathbb{E}[R_1]$, then with probability 1:

$$\lim_{K\to\infty} \frac{\sum_{k=1}^{K} R_k}{\sum_{k=1}^{K} T_k} = \frac{\mathbb{E}[R_1]}{\mathbb{E}[T_1]}$$

This important theorem is proven in just a few lines using the trick of multiplying and dividing by the same thing.

Proof: We have for each positive integer K:

$$\frac{\sum_{k=1}^{K} R_k}{\sum_{k=1}^{K} T_k} = \frac{\frac{1}{K}\sum_{k=1}^{K} R_k}{\frac{1}{K}\sum_{k=1}^{K} T_k}$$

Taking a limit as $K \to \infty$ and using the law of large numbers on the numerator and denominator on the right-hand-side thus yields:

$$\lim_{K\to\infty} \frac{\sum_{k=1}^{K} R_k}{\sum_{k=1}^{K} T_k} = \frac{\mathbb{E}[R_1]}{\mathbb{E}[T_1]} \quad \text{with probability 1}$$

□

One might argue that the relatively simple proof of this theorem is due to the fact that it takes advantage of the heavy lifting achieved by the proof of the law of large numbers. The proof of the law of large numbers is deep and requires several clever tricks. However, one can also argue that the *result* of the law of large numbers is intuitive. Hence, one can get immediate insight into renewal-reward processes by using the intuition for the law of large numbers together with the trick of multiplying and dividing by the same thing.

V. "THERE EXISTS" AND "FOR ALL"

Here is a problem designed to illustrate the difference between "there exists" and "for all." This may help students navigate through what they perceive as mathematical mumbo jumbo.

a) Let $\mathcal{A}$ be a subset of $\mathbb{R}^2$. Explain the difference between the following two statements:[3]

- Statement 1: There exists a vector $(x, y) \in \mathcal{A}$ such that $x + y = 0$.
- Statement 2: For all $(x, y) \in \mathcal{A}$, we have $x + y = 0$.

b) Give an example of a set $\mathcal{A}$ that satisfies statement 1, but not statement 2.

c) Suppose we want to show that statement 2 is true for a given set $\mathcal{A}$. Consider the proof structure below:

Proof: Fix any $(x, y) \in \mathcal{A}$. We want to show that ...[complete the sentence]. □

[3]Note that Statement 2 can be equivalently written "For any $(x, y) \in \mathcal{A}$, we have $x + y = 0$." This particular wording emphasizes that the best way of proving a given set $\mathcal{A}$ has the desired property is to write: "Fix any $(x, y) \in \mathcal{A}$. We want to show that..."