Pumping Lemma

Rebecca J Young

The Lemma

The Pumping Lemma is just a lot of words and symbols to communicate one (rather simple) idea:

> Any long-enough word that comes from a regular language must contain a loop.

Or, put another way: Every regular language has some number N, such that all words that are in that language that are longer than N, in their first N characters will contain a substring y which can be either removed or repeated to make more words in that language.

Why is this true?

Every regular language has some FA that recognises it. (I.e., for every regular language there exists some FA that accepts those words that are in the language; and rejects those words that are not.)

Every FA has a finite number of states. (Hence the 'F' for 'Finite'.)

Thus, given some regular language L, any word in L that has a length that is equal or greater than the number of states in the FA that recognises L *must* visit at least one state, S, more than once in its path from Start State to Final State. The characters read between visiting S the first time and visiting S the second time are part of a loop. Thus those characters make the substring y which can be removed (skip the loop) or repeated (loop the loop) to create a word that is also in L.

Note: This is trivially true for finite languages. In that case, there is no word that is of length $> N$ that is in the language, thus it is trivially true that all words of length $> N$ that are in the language contain a loop.... it just so happens that 'all words' are 'no words'.

How to write this formally?

Every regular language has some number N, such that all words that are in that language that are longer than N, will contain some substring y in their first N characters that can be either removed or repeated to make more words in that language.

- 🟪 There exists some number N such that: $\exists N$
- 🟩 For all words in the language that are longer than N: $\forall w : (|w| > N \land w \in L)$
- 🟥 There exists some way to break the word up such that a substring that occurs in the first N characters is isolated in y: $\exists x, y, z : (w = xyz) \land (y \neq \varepsilon) \land (|xy| \leq N)$
- 🟦 The substring can be removed or repeated to make more words in the language: $\forall i \geq 0 : xy^iz \in L$

Glue this together to get:

$$\exists N \forall w : (|w| > N \land w \in L) \rightarrow (\exists x, y, z : (w = xyz) \land (y \neq \varepsilon) \land (|xy| \leq N) \land (\forall i \geq 0 : xy^iz \in L))$$

How is this useful?

Every regular language satisfies the pumping lemma.

That is to say, **if** a language is regular, **then** the pumping lemma will be true for that language.

But it is **NOT THE CASE** that **if** the pumping lemma is true for a language **then** the language is regular.

YOUR TUTOR WHEN THEY ASK FOR A PROOF THAT A LANGUAGE IS REGULAR
AND YOU GIVE A PUMPING LEMMA PROOF
imgflip.com

If you are asked to prove that a language is regular, there are four valid approaches you may take:

- write a regular expression;
- draw an FA;
- draw an NFA;
- draw a GNFA.

Note that none of these options are 'use the pumping lemma'. Showing that the pumping lemma is satisfied proves *nothing* about your language.

So... how is the Pumping Lemma useful?

If a language is regular **then** it satisfies the Pumping Lemma.
Consequently: **If** a language does *not* satisfy the Pumping Lemma **then** it is *not* regular.

Pumping Lemma Proofs

The Pumping Lemma is a useful tool for proving that a given language is *not* regular.

The key insight is:

> If a language can make a long-enough word that does not contain a loop then that language cannot be regular.

The proof structure used in FIT2014 is a proof by contradiction. This structure is not strictly necessary, and you will see examples of proofs from different sources that do not utilise this structure. But if you are a FIT2014 student, follow the proof by contradiction steps.

Proof example 1: Prove that the language $a^m ba^n$ where $m < n$ is not regular.

Assumption: Let L be the language $a^m ba^n$ where $m < n$.
We assume, for the purpose of contradiction, that L is a regular language.

Discussion:
As L is regular, by assumption, by Kleene's Theorem there exists some FA that recognises L. Let N be the number of states in that FA.
As L is regular, by assumption, it satisfies the Pumping Lemma. That is, for all words $w \in L$ where $|w| > N$, there exists x, y, z such that:

1. $w = xyz$,
2. $y \neq \varepsilon$,
3. $|xy| \leq N$, and
4. for all $i \geq 0$, $xy^i z \in L$.

Let $w = a^N ba^{N+1}$. Note that $w \in L$ and $|w| > N$.

As $|xy| \leq N$ (rule 3), we reason that y must contain some number of as. That is, $y = a^\theta$ where $1 \leq \theta \leq N$.

However, when $i = 2$, $xy^i z = xyyz = a^{N+|y|} ba^{N+1}$. Thus we create the word $a^m ba^n$ where $m = N + |y|$ and $n = N + 1$. This means that $m \geq n$. Yet for a word to be in L, $m < n$. Thus $xy^2 z$ cannot be a word in L.

Contradiction: Thus L does not satisfy the Pumping Lemma, but as a regular language it must satisfy the Pumping Lemma.
This is a contradiction. We must conclude that it is not true that L is a regular language.

Conclusion: We have proven using contradiction that L is not a regular language.

Commentary 1: Prove that the language $a^m b a^n$ where $m < n$ is not regular.

Assumption: Let L be the language $a^m b a^n$ where $m < n$.

We assume, for the purpose of contradiction, that L is a regular language.

Discussion:

As L is regular, by assumption, by Kleene's Theorem there exists some FA that recognises L. Let N be the number of states in that FA.

As L is regular, by assumption, it satisfies the Pumping Lemma. That is, for all words $w \in L$ where $|w| > N$, there exists x, y, z such that:

1. $w = xyz$,
2. $y \neq \varepsilon$,
3. $|xy| \leq N$, and
4. for all $i \geq 0$, $xy^i z \in L$.

Let $w = a^N b a^{N+1}$. Note that $w \in L$ and $|w| > N$.

As $|xy| \leq N$ (rule 3), we reason that y must contain some number of as. That is, $y = a^\theta$ where $1 \leq \theta \leq N$.

However, when $i = 2$, $xy^i z = xyyz = a^{N+|y|} b a^{N+1}$. Thus we create the word $a^m b a^n$ where $m = N + |y|$ and $n = N + 1$. This means that $m \geq n$. Yet for a word to be in L, $m < n$. Thus $xy^2 z$ cannot be a word in L.

Contradiction: Thus L does not satisfy the Pumping Lemma, but as a regular language it must satisfy the Pumping Lemma.

This is a contradiction. We must conclude that it is not true that L is a regular language.

Conclusion: We have proven using contradiction that L is not a regular language.

Black/white text is general to any pumping lemma proof. Coloured text is specific to this proof. Coloured boxes refer to features of proof by contradiction.

Much of the discussion is just setting up the pumping lemma.

We choose a word where the first N characters are the same. This will minimise the number of cases we need to consider. We also choose a word that is one character away from 'breaking'. The chosen word just needs one extra a at the start to no longer be a word in the language.

We show that no matter which y is considered we can find some i that will break the word.

Proof example 2: Prove that the language of strictly increasing sequences of numbers in unary that contain three is not regular.

For a word to be in this language it must:

- be over the alphabet $\{1, ,\}$;
- express numbers using unary, where numbers are separated using ,;
- the numbers must be strictly increasing;
- the number three (111) must be present in the sequence.

For example:

- the word $1, 11, 111, 11111, 111111111$ is a word in this language;
- the word $11, 1111111, 111, 11111111$ is not a word in this language as it represents the sequence (2, 7, 3, 8) which is not strictly increasing;
- the word $1, 1111, 111111, 1111111$ is not a word in this language as it represents the sequence (1, 4, 6, 7) which does not include the number three.

Assumption: Let L be the language of strictly increasing sequences of numbers in unary that contain three. We assume, for the purpose of contradiction, that L is a regular language.

Discussion:
As L is regular, by assumption, by Kleene's Theorem there exists some FA that recognises L. Let N be the number of states in that FA.
As L is regular, by assumption, it satisfies the Pumping Lemma. That is, for all words $w \in L$ where $|w| > N$, there exists x, y, z such that:

1. $w = xyz$,
2. $y \neq \varepsilon$,
3. $|xy| \leq N$, and
4. for all $i \geq 0$, $xy^iz \in L$.

Let $w = 111, 1^N, 1^{N+1}$. Note that $w \in L$ and $|w| > N$.

As $|xy| \leq N$ (rule 3), we reason that there are three cases we need to consider:
Case 1: y contains some positive number of 1s from the representation of three.
Case 2: y contains some positive number of 1s from the representation of the number N.
Case 3: y contains the , and may contain some number of 1s from either side.

Case 1: If y contains only 1s taken from the representation of three, then when $i = 0$ the number three will be transformed into a smaller number, thus removing the number three from the sequence. Thus xy^0z cannot be a word in L. This is not a valid x, y, z.

Case 2: If y contains only 1s taken from the representation of N, then when $i = 2$ the number N will be transformed into the number $N + |y|$. As $N + |y| \geq N + 1$ the sequence will no longer be strictly increasing. Thus xy^2z cannot be a word in L. This is not a valid x, y, z.

Case 3: If y contains the , and m 1s taken from three and n 1s taken from N (where $0 \leq m \leq 3$ and $0 \leq n \leq N - 4$), then when $i = 3$ the number $m + n$ will be created twice in a row between 3 and N. As the same number will occur twice, the sequence will no longer be strictly increasing. Thus xy^3z cannot be a word in L. This is not a valid x, y, z.

After considering all possible cases, we conclude that we cannot find a valid x, y, z.

Contradiction: Thus L does not satisfy the Pumping Lemma, but as a regular language it must satisfy the Pumping Lemma.
This is a contradiction. We must conclude that it is not true that L is a regular language.

Conclusion: We have proven using contradiction that L is not a regular language.

Commentary 2: Prove that the language of strictly increasing sequences of numbers in unary that contain three is not regular.

Assumption: Let L be the language of strictly increasing sequences of numbers in unary that contain three.
We assume, for the purpose of contradiction, that L is a regular language.

Discussion:
As L is regular, by assumption, by Kleene's Theorem there exists some FA that recognises L. Let N be the number of states in that FA.
As L is regular, by assumption, it satisfies the Pumping Lemma. That is, for all words $w \in L$ where $|w| > N$, there exists x, y, z such that:

1. $w = xyz$,
2. $y \neq \varepsilon$,
3. $|xy| \leq N$, and
4. for all $i \geq 0$, $xy^iz \in L$.

Let $w = 111, 1^N, 1^{N+1}$. Note that $w \in L$ and $|w| > N$.

As our word requires that we include 111, we are unable to have the same character N times from the start of the word. We try to get to our grouping of N identical characters as early as possible.

As $|xy| \leq N$ (rule 3), we reason that there are three cases we need to consider:
Case 1: y contains some positive number of 1s from the representation of three.
Case 2: y contains some positive number of 1s from the representation of the number N.
Case 3: y contains the , and may contain some number of 1s from either side.

We need to consider multiple cases, as our y can fall anywhere in the first N characters.

Case 1: If y contains only 1s taken from the representation of three, then when $i = 0$ the number three will be transformed into a smaller number, thus removing the number three from the sequence. Thus xy^0z cannot be a word in L. This is not a valid x, y, z.

We can use a different i for each of our cases. We are not restricted to using the same i for every case.

Case 2: If y contains only 1s taken from the representation of N, then when $i = 2$ the number N will be transformed into the number $N + |y|$. As $N + |y| \geq N + 1$ the sequence will no longer be strictly increasing. Thus xy^2z cannot be a word in L. This is not a valid x, y, z.

Case 3: If y contains the , and m 1s taken from three and n 1s taken from N (where $0 \leq m \leq 3$ and $0 \leq n \leq N - 4$), then when $i = 3$ the number $m + n$ will be created twice in a row between 3 and N. As the same number will occur twice, the sequence will no longer be strictly increasing. Thus xy^3z cannot be a word in L. This is not a valid x, y, z.

After considering all possible cases, we conclude that we cannot find a valid x, y, z.

We cannot say that *no* x, y, z exists that satisfies the pumping lemma until we have tried *every* possible x, y, z.

Contradiction: Thus L does not satisfy the Pumping Lemma, but as a regular language it must satisfy the Pumping Lemma.
This is a contradiction. We must conclude that it is not true that L is a regular language.

Conclusion: We have proven using contradiction that L is not a regular language.