How to Negate Formulas

First-order logic gives a way both to rigorously define various terms and definitions and to help guide the structures of proofs of those definitions. If you ever want to prove a statement by contradiction, or simplify it to prove it by contrapositive, you will at some point need to take the negation of a first-order logic formula. While in many cases you can just "eyeball" the formula and figure out what the negation should be, for more complicated formulas it can be quite difficult to immediately spot the negation of the formula.

Fortunately, there's a nice, simple algorithm you can use to take the negation of any propositional or first-order logic formula. Even though first-order logic statements can look really dense and complicated, each formula is built purely from predicates, functions, $\wedge$, $\vee$, $\neg$, $\rightarrow$, $\leftrightarrow$, $\top$, $\bot$, $\forall$, and $\exists$. Consequently, to negate any first-order formula, you just have to know how to negate each of these constructs independently.

Below is a *very useful chart* showing how to simplify the negation of each of those constructs:

$\neg\neg\varphi$	$\neg(\varphi \wedge \psi)$	$\neg(\varphi \vee \psi)$	$\neg(\varphi \rightarrow \psi)$	$\neg(\varphi \leftrightarrow \psi)$
φ	$\varphi \rightarrow \neg\, \psi$	$\neg\varphi \wedge \neg\psi$	$\varphi \wedge \neg\psi$	$\varphi \leftrightarrow \neg\psi$

$\neg\top$	$\neg\bot$	$\neg\forall x.\ \varphi$	$\neg\exists x.\ \varphi$
$\bot$	$\top$	$\exists x.\ \neg\varphi$	$\forall x.\ \neg\varphi$

You can read this chart as follows. Anything of the form

$$\frac{X}{Y}$$

means "if you see X, simplify it by replacing it with Y." For example, the first entry says "if you see something of the form $\neg\neg\varphi$, you can simplify it by replacing it with φ. In this chart, I've used φ and ψ (the Greek letters phi and psi) as placeholders for more complicated formulas.

Let's try a simple example. Suppose we want to negate this formula:

$$\forall x.\ (\mathit{Puppy}(x) \rightarrow \mathit{Cute}(x))$$

We could try figuring out the negation of this formula by thinking about what it means, manipulating that meaning intuitively, then translating it back into logic. In fact, before moving on, why don't you take a few minutes and do just that? We can then reconvene on the next page and see if what you came up with agrees with what we ended up with.

Okay, so by now you should have a formula written down somewhere that you think is the negation of the above formula. I say "you should" because chances are that most of you didn't actually do that. So please – before we move on, go try this on your own first. Okay? Okay. ☺

Now, let's see how to do this mechanically. To begin with, let's put a negation at the front of the formula, like this:

$$\neg\forall x.\ (\mathit{Puppy}(x) \to \mathit{Cute}(x))$$

We'd like to push that negation as deep as possible. To do so, let's look at our table of negations. The formula that we have right here is a universally-quantified statement. If we look in our table, we see that we have this rule for negating universally-quantified statements:

$$\frac{\neg\forall x.\ \varphi}{\exists x.\ \neg\varphi}$$

In this particular example, our placeholder φ stands in for the statement $(\mathit{Puppy}(x) \to \mathit{Cute}(x))$. Therefore, we can simplify our formula by rewriting it as $\exists x.\ \neg\varphi$. This gives us the following:

$$\exists x.\ \neg(\mathit{Puppy}(x) \to \mathit{Cute}(x))$$

Let's keep pushing that negation deeper. The part of the formula that's negated is an implication: in this case, it's the implication $\mathit{Puppy}(x) \to \mathit{Cute}(x)$. How do we negate that? Well, consulting the table on the previous page, we see this rule:

$$\frac{\neg(\varphi \to \psi)}{\varphi \wedge \neg\psi}$$

Here, φ is the statement $\mathit{Puppy}(x)$ and ψ is the statement $\mathit{Cute}(x)$. This means that $\varphi \wedge \neg\psi$ is the formula $\mathit{Puppy}(x) \wedge \neg\mathit{Cute}(x)$. Replacing $\neg(\mathit{Puppy}(x) \to \mathit{Cute}(x))$ in our formula with this yields the following:

$$\exists x.\ (\mathit{Puppy}(x) \wedge \neg\mathit{Cute}(x))$$

At this point, we can't simplify any further – the only negation that exists is directly applied to a predicate. This means that we're done – what we have above is the correct negation of the original statement.

What does this statement say in plain English? We can read it as "there is some puppy that is not cute." Our original statement says "every puppy is cute." Those two statements are indeed opposites of one another, so it seems like we got the negation right!

How does that compare to the negation you came up with earlier? Does it match? Did you end up with something different? If your answer is different from ours, go talk to your problem set group about it and get their thoughts. If you're unsure, go and ask us on Piazza or in office hours,

Let's do another example, just for fun. Let's try this statement:

$$(\forall x.\ \mathit{Happy}(x)) \to (\exists x.\ \mathit{Happy}(x))$$

What's the negation of this statement? We'll derive it on the next page. Like last time, though, you should take a minute to try to figure this out on your own first. Write down your thoughts, then rejoin us on the next page.

Did you actually try? You really should – this is a great exercise! If not, oh well, your loss. ☺

This statement is a bit trickier than the previous one because, if you look closely, you'll notice that the entire statement isn't quantified. Rather, the statement is an implication consisting of one quantified statement implying another quantified statement. Therefore, if we want to negate the entire statement, we'll start with something like this:

$$\neg((\forall x.\ \mathit{Happy}(x)) \rightarrow (\exists x.\ \mathit{Happy}(x)))$$

How do we simplify this? Well, as we said before, this statement is actually a giant implication being negated. Going back to our rule table, we see this:

$$\frac{\neg(\varphi \rightarrow \psi)}{\varphi \wedge \neg\psi}$$

So in order to simplify the negation, we're going to switch the → to a ∧ and push the negation to the consequent. This gives us the following:

$$(\forall x.\ \mathit{Happy}(x)) \wedge \neg(\exists x.\ \mathit{Happy}(x))$$

Now, we've got the negation of an existential statement. Our rule table tells us the following:

$$\frac{\neg\exists x.\ \varphi}{\forall x.\ \neg\varphi}$$

Applying this rule, we get the following:

$$(\forall x.\ \mathit{Happy}(x)) \wedge (\forall x.\ \neg\mathit{Happy}(x))$$

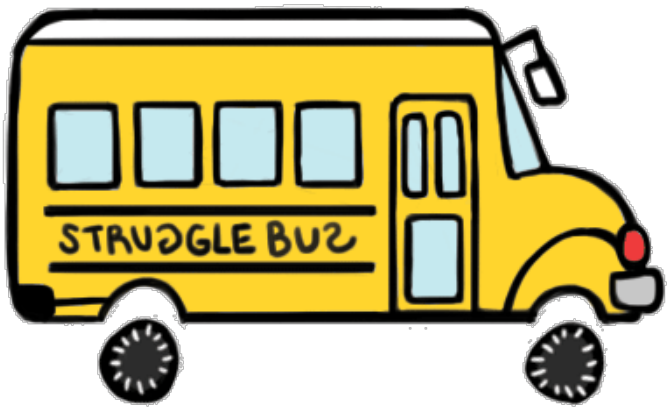
How does that look compared to what you came up with initially? Do they agree? Did you find an alternative negation? If what you have doesn't match what we came up with above, it doesn't mean it's wrong. After all, with a bit more thought, you can simplify this a bit further, though doing so will require you to understand at a deep level what this statement actually says.

Before concluding, let's do one final example, one that's a lot harder than the others. Suppose you want to negate this formula:

$$\forall S.\ \forall T.\ (\mathit{Set}(S) \wedge \mathit{Set}(T) \rightarrow (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

Before reading how we negated this statement, go and try negating it on your own. This one's tough, so go one step at a time. See what you come up with. Then, read onward to see what we did.

Seriously, don't read the next part until you've actually tried negating this statement on your own. You should hop on the struggle bus and make a good effort before seeing our solution, since it's great practice.



Okay – let's see how to do this.

We begin by putting a negation in the front of the whole formula, as shown here:

$$\neg\forall S.\ \forall T.\ (Set(S) \wedge Set(T) \to (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

Here, we have the negation of a universally-quantified statement. We have this rule for negating universally-quantified statements:

$$\frac{\neg\forall x.\ \varphi}{\exists x.\ \neg\varphi}$$

Applying that rule gives us the following:

$$\exists S.\ \neg\forall T.\ (Set(S) \wedge Set(T) \to (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

We now have another negation in front of a universal quantifier. Applying the same rule tells us that we can simplify this formula down to

$$\exists S.\ \exists T.\ \neg(Set(S) \wedge Set(T) \to (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

What do we do now? Notice that the part of the formula that's currently being negated is this part here:

$$\neg(Set(S) \wedge Set(T) \to (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T))$$

This looks really complicated. How do we negate it? Well, let's start by figuring out the structure of the formula. We've got an $\wedge$, an $\to$, two $\leftrightarrow$'s, and a $\forall$. If we add in parentheses, we can see that the overall structure of this part of the formula is

$$\neg((Set(S) \wedge Set(T)) \to (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T))$$

Notice that the overall shape of this formula is

$$\neg(\varphi \to \psi)$$

Where φ is $(Set(S) \wedge Set(T))$ and ψ is $(S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T))$. (Do you see why?) We have a rule saying how we can simplify this:

$$\frac{\neg(\varphi \to \psi)}{\varphi \wedge \neg\psi}$$

This means that we can rewrite $\neg(\varphi \to \psi)$ as $\varphi \wedge \neg\psi$. In our case, doing so takes us from

$$\exists S.\ \exists T.\ \neg((Set(S) \wedge Set(T)) \to (S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

to

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge \neg(S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

We're making more progress! We now have this part of the formula to simplify:

$$\neg(S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T))$$

Fortunately, we have a rule for dealing with it:

$$\frac{\neg(\varphi \leftrightarrow \psi)}{\varphi \leftrightarrow \neg\psi}$$

Applying this rule lets us simplify

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge \neg(S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

to

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge (S = T \leftrightarrow \neg\forall x.\ (x \in S \leftrightarrow x \in T)))$$

Applying our earlier rule about negating universal quantifiers takes us to

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge (S = T \leftrightarrow \exists x.\ \neg(x \in S \leftrightarrow x \in T)))$$

And, finally, applying our rule about negating biconditionals takes us to

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge (S = T \leftrightarrow \exists x.\ (x \in S \leftrightarrow x \notin T)))$$

And we're done! Notice that we've pushed the negation as deep into the formula as possible. No further simplifications are possible, so we've got the negation of our original statement.

It turns out that there's a slightly easier way to negate this statement. Let's back up to this point:

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge \neg(S = T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

Here, we have the negation of a biconditional statement. The rule in our table says that we negate statements like these as follows:

$$\frac{\neg(\varphi \leftrightarrow \psi)}{\varphi \leftrightarrow \neg\psi}$$

However, this isn't the only legal way to negate a biconditional. We could also do the following:

$$\frac{\neg(\varphi \leftrightarrow \psi)}{\neg\varphi \leftrightarrow \psi}$$

Take a few minutes to convince yourself that this works. Check out the truth table tool – what does it say about the truth tables here? Intuitively, does this make sense?

Applying this rule here to simplify the negated biconditional would give us the following:

$$\exists S.\ \exists T.\ (Set(S) \wedge Set(T) \wedge (S \neq T \leftrightarrow \forall x.\ (x \in S \leftrightarrow x \in T)))$$

Notice that at this point there's no further simplification possible. Since we never negated the innermost universal quantifier, there was no need to push the negation across that quantifier. Nifty!

Notice that every step of the way, we just applied one minor simplification at a time by figuring out what the general structure of the formula being negated was and, from there, applying one simple rule at a time to simplify the formula. There was never any question about “what does this formula mean?” or “how would we express the opposite idea of this formula?” - it was a purely mechanical process, the sort of thing that we computer scientists tend to really like.

Of course, it's good to think about what that initial formula means and what this negation means, but we'll leave that as an exercise. ☺