FIRST-ORDER LOGIC

1. First-order formulas

First-order logic is an extension of propositional logic allowing us to express statements *about elements*, instead of just pure statements. Here is an example of a first-order formula:

$$\forall x\,(0 \le x \to \exists y\,((0 \le y) \wedge (y \cdot y = x)))$$

There are several new syntactic constructions available, compared to propositional logic:

- There are two types of expressions: formulas like $y \ge 0$, which express statements (true or false), as well as **terms** like $y \cdot y$, which denote elements rather than statements.
- Both formulas and terms may depend on **variables** like x, y. In other words, a formula $y \cdot y = x$ represents not a single truth value but rather a *relation* (binary in this case).
- **Quantifiers** like $\forall, \exists$ allow us to *bind* variables in formulas: for example, the formula $\exists y\,((y \ge 0) \wedge (y \cdot y = x))$ no longer depends on y.
- There are some atomic symbols like $\le$, called **relation symbols**, to be specified in the alphabet $\mathcal{A}$, that can be used to combine terms into formulas. ($=$ can also be regarded as a binary relation symbol, although it plays a rather special role.)
- There are some other atomic symbols like $\cdot$, called **function symbols** (or **operation symbols**), also specified by $\mathcal{A}$, that can be used to combine terms into other terms. 0 above can also be regarded as a (0-ary) function symbol.

The formal definition is as follows.

Definition 1.1. A **first-order signature** is an alphabet $\mathcal{A}$ together with, for each $P \in \mathcal{A}$, two additional pieces of data:

- a classification of P as either a **relation symbol** or a **function symbol**;
- an **arity** $n \in \mathbb{N}$; we call P **n-ary** (or **binary** when $n = 2$, **unary** when $n = 1$, etc.).

A 0-ary (or "nullary") function symbol is also called a **constant symbol**. We write

$$\begin{aligned}
\mathcal{A}_{\text{rel}} &:= \{\text{relation symbols in } \mathcal{A}\} \subseteq \mathcal{A},\\
\mathcal{A}_{\text{fun}} &:= \{\text{function symbols in } \mathcal{A}\} \subseteq \mathcal{A},\\
\mathcal{A}^n_{\text{rel}} &:= \{n\text{-ary relation symbols}\} \subseteq \mathcal{A}_{\text{rel}},\\
\mathcal{A}^n_{\text{fun}} &:= \{n\text{-ary function symbols}\} \subseteq \mathcal{A}_{\text{fun}}.
\end{aligned}$$

Thus, formally, a first-order signature consists of a set $\mathcal{A}$ equipped with a partition

$$\begin{aligned}
\mathcal{A} &= \textstyle\bigsqcup_{n\in\mathbb{N}} \mathcal{A}^n_{\text{rel}} \sqcup \bigsqcup_{n\in\mathbb{N}} \mathcal{A}^n_{\text{fun}}\\
&= \mathcal{A}^0_{\text{rel}} \sqcup \mathcal{A}^1_{\text{rel}} \sqcup \cdots \sqcup \mathcal{A}^0_{\text{fun}} \sqcup \mathcal{A}^1_{\text{fun}} \sqcup \cdots .
\end{aligned}$$

However, in practice, we usually just list out the elements of the alphabet $\mathcal{A}$, and then say in words what type of symbol each element is; for familiar symbols, like $\le, +$, we usually take them to be of the familiar type and arity.

Example 1.2. The **signature of graphs** is $\mathcal{A}_{\mathsf{graph}} := \{E\}$, where E is a binary relation symbol.

Example 1.3. The **signature of posets** is $\mathcal{A}_{\mathsf{poset}} := \{\le\}$, where $\le$ is a binary relation symbol. Note that this is identical to the signature of graphs, except for the symbol we chose to use.

Remark 1.4. The preceding two examples indicate an important point: *a signature can only specify what the relations/operations are; it cannot specify how they behave.* For example, $\mathcal{A}_{\mathsf{poset}}$ does not specify transitivity of $\leq$ in any way. In order to specify axioms that the relations/operations have to obey, we need a first-order theory (see Section 2.3 below).

Example 1.5. The **signature of fields** is $\mathcal{A}_{\mathsf{field}} := \{+, 0, -, \cdot, 1\}$ where the symbols are, respectively, (2, 0, 1, 2, 0)-ary function symbols (so $0, 1$ are constant symbols).

(We do not include a symbol for $/$, because division is not an everywhere-defined operation; we can only require that nonzero elements in a field must have a multiplicative inverse, via an axiom in the *theory* of fields (see Example 2.25 below). Thus, this signature could just as well be called the signature of *rings* $\mathcal{A}_{\mathsf{ring}}$, illustrating again the preceding remark.)

Example 1.6. The **signature of ordered fields** is $\mathcal{A}_{\mathsf{ordfield}} := \mathcal{A}_{\mathsf{field}} \cup \mathcal{A}_{\mathsf{poset}} = \{+, 0, -, \cdot, 1, \leq\}$.

Example 1.7. The **signature of ($\mathbb{R}$-)vector spaces** is $\mathcal{A}_{\mathsf{vec}} := \{+, 0, -, a\cdot \mid a \in \mathbb{R}\}$, where $+, 0, -$ are as in the signature of fields above, while for each $a \in \mathbb{R}$, $a\cdot$ is a single *unary* function symbol (referring to scalar multiplication by a). So $\mathcal{A}_{\mathsf{vec}}$ is an infinite (indeed uncountable) signature.

(It would not make sense to treat $\cdot$ as a binary function symbol if we want to use this signature to describe vector spaces, since scalar multiplication does not take two vectors in a vector space V to another vector.)

Definition 1.8. Let $\mathcal{A}$ be a first-order signature. Fix also another alphabet $\mathcal{V}$, whose elements we call **variables**. The **$\mathcal{A}$-terms with variables from $\mathcal{V}$** are constructed inductively as follows:

- Every $x \in \mathcal{V}$ is an $\mathcal{A}$-term.
- If $f \in \mathcal{A}^n_{\mathrm{fun}}$ is an n-ary function symbol, and $t_1, \ldots, t_n$ are terms, then so is $f(t_1, \ldots, t_n)$.

The **(first-order) $\mathcal{A}$-formulas with variables from $\mathcal{V}$** are constructed inductively as follows:

- If $R \in \mathcal{A}^n_{\mathrm{rel}}$ is an n-ary relation symbol, or the symbol $=$ when $n = 2$, and $t_1, \ldots, t_n$ are $\mathcal{A}$-terms, then $R(t_1, \ldots, t_n)$ is an $\mathcal{A}$-formula, called an **atomic formula**.[1]
- If ϕ, ψ are $\mathcal{A}$-formulas, then $\phi \wedge \psi, \phi \vee \psi, \neg\phi$ are $\mathcal{A}$-formulas.
- $\top, \bot$ are $\mathcal{A}$-formulas.
- If ϕ is an $\mathcal{A}$-formula, and $x \in \mathcal{V}$ is a variable, then $\exists x\, \phi$ is an $\mathcal{A}$-formula.

We continue to use the abbreviations $\rightarrow, \leftrightarrow$ as in propositional logic, as well as

$$\forall x\, \phi := \neg\exists x\, \neg\phi.$$

(The reason for regarding $\forall$ as an abbreviation, rather than $\exists$, is similar to why we chose to regard $\rightarrow$ as an abbreviation in propositional logic, but not $\phi \vee \psi := \neg(\neg\phi \wedge \neg\psi)$, say: we will use them to illustrate different aspects of our proof system for first-order logic. Indeed, there is a sense in which $\forall$ is analogous to $\rightarrow$ and $\exists$ to $\vee$; see Remark 4.18 and Example 4.27 below.)

Example 1.9. Let $x, y \in \mathcal{V}$ be variables. The following is an $\mathcal{A}_{\mathsf{ordfield}}$-term:

$$+(1, \cdot(x, y))$$

When we are dealing with signatures consisting of familiar symbols like $+, \cdot$, we will write terms and formulas in the familiar way; e.g., the above term would usually be written

$$1 + x \cdot y.$$

Likewise, the $\mathcal{A}_{\mathsf{ordfield}}$-formula given at the beginning of this section is a more familiar way of writing

$$\forall x\, (\leq(0, x) \rightarrow \exists y\, (\leq(0, y) \wedge =(\cdot(y, y), x))).$$

[1] There is an abuse of notation going on here: for two terms s, t, "$s = t$" may denote *either* the "meta" assertion that these two terms are the same (as expressions), *or* the atomic formula $s = t$! Some logic books therefore use a different symbol (like $\equiv$) for the equality formula. We will instead depend on context to disambiguate.

Example 1.10. The following are *not* $\mathcal{A}_{\mathsf{ordfield}}$-formulas:

$\leq(x, y, z)$	($\leq$ is binary, not ternary)
$\forall x\,(x + y \cdot y)$	($\forall x$ must be followed by a formula, not a term)
$\forall x\,(\bot \leq x \cdot x)$	(the LHS of $\leq$ must be a term, not a formula)
$\forall x\,(\sqrt{x} \cdot \sqrt{x} = x)$	(no $\sqrt{\ }$ symbol in $\mathcal{A}_{\mathsf{ordfield}}$)
$\forall x\,(2 + x = x + 2)$	(no 2 symbol in $\mathcal{A}_{\mathsf{ordfield}}$)

However, we might treat the last formula as an abbreviation for

$$\forall x\,((1 + 1) + x = x + (1 + 1)).$$

On the other hand, the following *are* $\mathcal{A}_{\mathsf{ordfield}}$-formulas:

$0 = 1$	(will be interpreted as false)
$\exists x\,\top$	(will be interpreted as "the model is nonempty")
$0 \leq x \to \forall x\,\exists x\,(x \leq 0)$	(nothing in definition of formula prevents variable clashes)

1.1. **Free and bound variables.** The last example above shows that in order to interpret formulas correctly, it is important to pay attention to which variables occur underneath quantifiers.

An occurrence of a variable underneath a quantifier in a formula is called **bound**; a **free variable** in a formula is a variable which occurs non-bound (at least once). Since terms do not contain quantifiers, all variables in terms are considered free. Formally, we define the **set of free variables** $\mathrm{FV}(t), \mathrm{FV}(\phi)$ of a term t or formula ϕ inductively as follows:

$$\begin{aligned}
\mathrm{FV}(x) &:= \{x\},\\
\mathrm{FV}(f(t_1, \ldots, t_n)) &:= \mathrm{FV}(t_1) \cup \cdots \cup \mathrm{FV}(t_n) \quad \text{for } f \in \mathcal{A}^n_{\mathrm{fun}} \text{ and } t_1, \ldots, t_n \in \mathcal{L}_{\mathrm{term}}(\mathcal{A}),\\
\mathrm{FV}(R(t_1, \ldots, t_n)) &:= \mathrm{FV}(t_1) \cup \cdots \cup \mathrm{FV}(t_n) \quad \text{for } R \in \mathcal{A}^n_{\mathrm{rel}} \text{ (or } R = =) \text{ and } t_1, \ldots, t_n \in \mathcal{L}_{\mathrm{term}}(\mathcal{A}),\\
\mathrm{FV}(\phi \wedge \psi) &:= \mathrm{FV}(\phi \vee \psi) := \mathrm{FV}(\phi) \cup \mathrm{FV}(\psi),\\
\mathrm{FV}(\neg\phi) &:= \mathrm{FV}(\phi),\\
\mathrm{FV}(\top) &:= \mathrm{FV}(\bot) := \varnothing,\\
\mathrm{FV}(\exists x\,\phi) &:= \mathrm{FV}(\phi) \setminus \{x\}.
\end{aligned}$$

(Compare with the definition of the set $\mathrm{AT}(\phi)$ of atomic formulas in Example 1.6 in the notes on propositional logic.) Recalling that we regard $\to$, $\leftrightarrow$, and $\forall$ as abbreviations, we also have

$$\begin{aligned}
\mathrm{FV}(\phi \to \psi) &= \mathrm{FV}(\neg\phi \vee \psi) = \mathrm{FV}(\phi) \cup \mathrm{FV}(\psi),\\
\mathrm{FV}(\phi \leftrightarrow \psi) &= \mathrm{FV}(\phi) \cup \mathrm{FV}(\psi),\\
\mathrm{FV}(\forall x\,\phi) &= \mathrm{FV}(\neg\exists x\,\neg\phi) = \mathrm{FV}(\phi) \setminus \{x\}.
\end{aligned}$$

Example 1.11. To compute the free variables of the $\mathcal{A}_{\mathsf{ordfield}}$-formula from the beginning of these notes:

$$\underbrace{\forall x\,(\overbrace{0 \leq x}^{\mathrm{FV}=\{x\}} \to \underbrace{\exists y\,(\overbrace{(0 \leq y)}^{\mathrm{FV}=\{y\}} \wedge \overbrace{(y \cdot y = x)}^{\mathrm{FV}=\{x,y\}})}_{\mathrm{FV}=\{y\}\cup\{x,y\}=\{x,y\}})}_{\mathrm{FV}=\{x,y\}\setminus\{y\}=\{x\}}}_{\mathrm{FV}=\{x\}\cup\{x\}=\{x\}}}_{\mathrm{FV}=\{x\}\setminus\{x\}=\varnothing}$$

Exercise 1.12. Compute the free variables of the following formulas:

(a) $0 \leq x \to \forall x\, \exists x\, (x \leq 0)$
(b) $(\forall x\, (x \leq x \cdot y)) \vee (\exists y\, (x \cdot y \leq y))$
(c) $\forall x\, ((\forall y\, (x \leq y \cdot z)) \to \exists x\, (x + y = z))$

A term or formula is called **closed** if it has no free variables; these denote a particular element or truth value, which does not depend on the values of any variables. A closed formula is also called a **sentence**, the above being an example. Non-closed formulas/terms are sometimes called **open**.[2]

The set of *free* variables in a formula is much more important than what *all* the variables are, since bound variables may be changed without affecting the meaning of the formula: for example,

$$\exists y\, (x + y = 0) \text{ vs. } \exists z\, (x + z = 0)$$

should always have the same meaning. (Of course, as always, these two formulas are not quite *equal*; but we will introduce the notion of "α-equivalence" in Section 3.2 below in order to identify them.) For this reason, from now on, we will never mention the set $\mathcal{V}$ from which all variables are drawn; we will only ever care what the free variables of a formula are. For any set of variables X, we write

$$\mathcal{L}^X_{\text{term}}(\mathcal{A}) := \{\mathcal{A}\text{-terms } t \mid \text{FV}(t) \subseteq X\},$$
$$\mathcal{L}^X_{\text{form}}(\mathcal{A}) := \{\mathcal{A}\text{-formulas } \phi \mid \text{FV}(\phi) \subseteq X\},$$

and call these the set of t, ϕ respectively with **free variables from** X.

Remark 1.13. It is important to note that when we say ϕ has free variables from X, we do not actually require each $x \in X$ to occur in ϕ; we only care that no other variables can occur free in ϕ. This is usually more important than knowing which free variables actually do occur: for example, as long as ϕ has free variables from $\{x, y\}$, then $\forall x\, \exists y\, \phi$ will be a sentence.

2. First-order semantics

2.1. **Structures.** Let $\mathcal{A}$ be a first-order signature. An **$\mathcal{A}$-structure** $\mathcal{M}$ consists of:

- an **underlying set** (also called **domain** or **universe**), denoted M or $|\mathcal{M}|$;
- for each n-ary relation symbol $R \in \mathcal{A}^n_{\text{rel}}$, an n-ary relation $R^{\mathcal{M}} = \mathcal{M}(R)$ on M;
- for each n-ary function symbol $f \in \mathcal{A}^n_{\text{fun}}$, an n-ary function $f^{\mathcal{M}} = \mathcal{M}(f) : M^n \to M$.

We call $R^{\mathcal{M}}, f^{\mathcal{M}}$ the **interpretation of** R, f **in** $\mathcal{M}$.

Here M^n denotes the n-fold Cartesian product

$$M^n := \underbrace{M \times \cdots \times M}_{n} = \{(a_1, \ldots, a_n) \mid a_1, \ldots, a_n \in M\}.$$

When $n = 1$, we usually identify M^1 with M, so that instead of writing 1-tuples (a) for $a \in M$ we may simply write a. When $n = 0$, $M^0 = \{()\}$ is a one-element set consisting of the empty tuple $()$. An 0-ary function $f : M^0 \to M$ thus consists of simply an element $f(()) \in M$; we usually identify f with $f(())$. Thus, each constant symbol $c \in \mathcal{A}^0_{\text{fun}}$ is interpreted as an element $c^{\mathcal{M}} \in M$.

An **n-ary relation** R on M may be represented in multiple ways:

- R may be thought of as a subset $R \subseteq M^n$, namely the set of all n-tuples at which R holds. For example, the equality relation on M is represented as

$$\begin{aligned}(=_M) &= \{(a, b) \in M^2 \mid a = b\} \subseteq M^2 \\ &= \{(a, a) \mid a \in M\}.\end{aligned}$$

  To say that R **holds at a tuple** $\vec{a} = (a_1, \ldots, a_n)$ then means that $\vec{a} \in R$.

[2]This terminology has nothing to do with "closed" and "open" sets in topology.

- R may also be thought of as a function $R : M^n \to \{0,1\}$, which specifies whether or not R holds at each n-tuple. For example, the equality relation on M is represented as

$$(=_M) : M^2 \longrightarrow \{0,1\}$$
$$(a,b) \longmapsto \begin{cases} 1 & \text{if } a = b, \\ 0 & \text{if } a \neq b. \end{cases}$$

  To say that R **holds at** $\vec{a}$ then means that $R(\vec{a}) = 1$.

Given a subset $R \subseteq M^n$, the corresponding function to $\{0,1\}$ is its **indicator function** (or **characteristic function**)

$$\chi_R : M^n \longrightarrow \{0,1\}$$
$$\vec{a} \longmapsto \begin{cases} 1 & \text{if } \vec{a} \in R, \\ 0 & \text{otherwise.} \end{cases}$$

Conversely, given a function $S : M^n \to \{0,1\}$, the corresponding subset is the preimage

$$S^{-1}(1) = \{\vec{a} \in M^n \mid S(\vec{a}) = 1\}.$$

The operations $R \mapsto \chi_R$ and $S \mapsto S^{-1}(1)$ are inverse bijections between the set of all subsets of M^n and the set of all functions $M^n \to \{0,1\}$; this is why we may think of either as representing n-ary relations on M. We will find it convenient to use both of these representations. Therefore, we adopt the following abuse of notation:

Convention 2.1. We identify relations represented as sets of tuples with their indicator functions. Thus, to say that an n-ary relation R holds at a tuple $\vec{a} \in M^n$ means either of

$$R(\vec{a}) = 1 \iff \vec{a} \in R.$$

When $n = 2$, we also sometimes adopt the traditional "infix" notation, as in $x \leq y$:

$$R(a,b) = 1 \iff (a,b) \in R \iff :\ a\ R\ b.$$

Note also that a 0-ary relation R is (a function $R : M^0 = \{()\} \to \{0,1\}$, hence) equivalently just a truth value, which is how we usually think of it.

Example 2.2. We have a $\mathcal{A}_{\mathsf{ordfield}}$-structure $\mathcal{R}$ with underlying set $\mathbb{R}$ and each symbol in $\mathcal{A}_{\mathsf{ordfield}}$ interpreted as the usual operation or relation of that name, e.g., $+^{\mathcal{R}} : \mathbb{R}^2 \to \mathbb{R}$ is the binary addition function. Similarly, we have a $\mathcal{A}_{\mathsf{ordfield}}$-structure $\mathcal{Q}$ consisting of $\mathbb{Q}$ and the usual interpretations.

Example 2.3. We have a $\mathcal{A}_{\mathsf{ordfield}}$-structure $\mathcal{M}$ with underlying set $\mathbb{R}$ and

$$+^{\mathcal{M}} := \text{usual } +,$$
$$0^{\mathcal{M}} := \text{usual } 0,$$
$$-^{\mathcal{M}} := \text{usual sin},$$
$$\cdot^{\mathcal{M}} := \text{usual } +,$$
$$1^{\mathcal{M}} := \text{usual } \pi,$$
$$\leq^{\mathcal{M}} := \text{usual } =.$$

(Nothing in the definition of $\mathcal{A}_{\mathsf{ordfield}}$-structure says that the field axioms like commutativity, etc., have to hold; this will be enforced by the first-order *theory* of fields, see Example 2.25.)

Example 2.4. Similarly, a $\mathcal{A}_{\mathsf{poset}}$-structure $\mathcal{M}$ is a set M equipped with an *arbitrary* binary relation $\leq^{\mathcal{M}} \subseteq M^2$ (not yet required to be a partial order).

Example 2.5. For $\mathcal{A} = \varnothing$, an $\mathcal{A}$-structure is just a set.

2.2. **Interpretation of terms and formulas.** Let $\mathcal{M}$ be an $\mathcal{A}$-structure. In order to interpret a term or formula in $\mathcal{M}$, we need to know what values are assigned to its free variables; in other words, the interpretation will be a *function* defined on the set of all variable assignments (to either M or $\{0,1\}$, depending on whether we have a term or a formula).

By a **variable assignment** in M, we just mean a function $\alpha : X \to M$ from some set of variables X. The set of all X-indexed variable assignments is thus M^X, the set of all functions from X to M. Note that we can also think of $\alpha : X \to M$ as an "X-ary tuple" of elements of M, namely $(\alpha(x))_{x\in X}$; this allows us to think of the interpretation of terms and formulas as generalizing the interpretation of function and relation symbols in $\mathcal{A}$ (which yield functions $M^n \to M$ or $M^n \to \{0,1\}$).

Definition 2.6. We begin with the interpretation of terms. For each $\mathcal{A}$-term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ with free variables from some set X, we will define by induction on t a function

$$t^{\mathcal{M}}_X = \mathcal{M}_X(t) : M^X \longrightarrow M,$$

called the **interpretation of t in $\mathcal{M}$**, which maps each variable assignment $\alpha \in M^X$ to an element $t^{\mathcal{M}}_X(\alpha) \in M$, called the **interpretation of t in $\mathcal{M}$ under the variable assignment α**:

- For a single variable $x \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ with free variables from X, this means $x \in X$; we define
$$x^{\mathcal{M}}_X(\alpha) := \alpha(x).$$
- For a term $f(t_1,\dots,t_n) \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ where $f \in \mathcal{A}^n_{\text{fun}}$ and $t_1,\dots,t_n \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we define
$$f(t_1,\dots,t_n)^{\mathcal{M}}_X(\alpha) := f^{\mathcal{M}}((t_1)^{\mathcal{M}}_X(\alpha),\dots,(t_n)^{\mathcal{M}}_X(\alpha))$$
(recall that $f^{\mathcal{M}} : M^n \to M$).

Example 2.7. In the $\mathcal{A}_{\mathsf{ordfield}}$-structure $\mathcal{M} = \mathbb{R}$ with weird operations from Example 2.3,

$$\begin{aligned}((-1)\cdot(x+y))^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5) &= (-1)^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5) \cdot^{\mathcal{M}} (x+y)^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5)\\ &= (-^{\mathcal{M}}1^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5)) + (3 +^{\mathcal{M}} 5)\\ &= \sin(\pi) + (3+5) = 8.\end{aligned}$$

Exercise 2.8. Verify that this is the same as $((-1)\cdot x + (-1)\cdot y)^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5)$.

Is it also the same as $((-x)+(-y))^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5)$?

Remark 2.9. You will probably have noticed how much redundant information we had to keep writing in the above example. Strictly speaking, all of this information is necessary: the subscript X on $t^{\mathcal{M}}_X$ is needed because we can always regard t as having more free variables, so that for example, $(x+y)^{\mathcal{M}}_{\{x,y\}} : M^{\{x,y\}} \to M$ and $(x+y)^{\mathcal{M}}_{\{x,y,z\}} : M^{\{x,y,z\}} \to M$ are two completely different functions with different domains. However, once we plug in α to $t^{\mathcal{M}}_X(\alpha)$, we can tell what the X is based on the domain of α; for this reason, we will often omit the X, so that we can write, e.g.,

$$(x+y)^{\mathcal{M}}(x\mapsto 3,\, y\mapsto 5) := (x+y)^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5).$$

The other major source of redundancy was how we had to keep writing the entire variable assignment, even when those variables stopped appearing, e.g., in $(-1)^{\mathcal{M}}_{\{x,y\}}(x\mapsto 3,\, y\mapsto 5)$. Again, this is necessary *a priori*, since the definition of $t^{\mathcal{M}}_X(\alpha)$ could potentially depend on all of α (and for the interpretation of formulas below, it is much less obvious that it doesn't, due to the $\exists$ case). Fortunately, you will prove on HW5 that the interpretation indeed remains unchanged when you drop variables which aren't free in the term/formula being interpreted:

Proposition 2.10 (HW5)**.** Let $\alpha : Y \to M$ be a variable assignment and $X \subseteq Y$.

(a) If a term t only has free variables from X, then $t^{\mathcal{M}}_X(\alpha|X) = t^{\mathcal{M}}_Y(\alpha)$.

(b) If a formula ϕ only has free variables from X, then $\phi^{\mathcal{M}}_X(\alpha|X) = \phi^{\mathcal{M}}_Y(\alpha)$.

Definition 2.11. When interpreting formulas, in the $\exists$ case, because the subformula has one more free variable, we will need to extend our given variable assignment to include that extra variable. We therefore introduce the following general notation: for a function $\alpha : X \to M$, a variable x (which may or may not be in X), and an element $a \in M$,

$$\alpha\langle x \mapsto a\rangle : X \cup \{x\} \longrightarrow M$$
$$y \longmapsto \begin{cases} a & \text{if } y = x, \\ \alpha(y) & \text{if } y \in X \setminus \{x\}. \end{cases}$$

In other words, we add the assignment $x \mapsto a$ to α, replacing any previous value of $\alpha(x)$.

Definition 2.12. We now give the interpretation of formulas. For an $\mathcal{A}$-formula $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ with free variables from X, its **interpretation** $\phi^{\mathcal{M}}_X = \mathcal{M}_X(\phi)$ **in** $\mathcal{M}$ will be an "X-ary relation" on M, hence by Convention 2.1 may be represented either as a set of "X-ary tuples"

$$\phi^{\mathcal{M}}_X = \mathcal{M}_X(\phi) \subseteq M^X$$

or its indicator function

$$\phi^{\mathcal{M}}_X = \mathcal{M}_X(\phi) : M^X \longrightarrow \{0, 1\}.$$

If $\phi^{\mathcal{M}}_X(\alpha) = 1$, i.e., $\alpha \in \phi^{\mathcal{M}}_X$, then we say that $\mathcal{M}$ **satisfies** ϕ **under** α, also denoted

$$\mathcal{M} \models_\alpha \phi \ :\Longleftrightarrow\ \phi^{\mathcal{M}}_X(\alpha) = 1 \iff \alpha \in \phi^{\mathcal{M}}_X.$$

We will give the inductive definition using all three of these equivalent notations at once.

- For $\phi = R(t_1, \ldots, t_n) \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ where $R \in \mathcal{A}^n_{\text{rel}}$ and $t_1, \ldots, t_n \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, we define
$$R(t_1, \ldots, t_n)^{\mathcal{M}}_X(\alpha) := R^{\mathcal{M}}((t_1)^{\mathcal{M}}_X(\alpha), \ldots, (t_n)^{\mathcal{M}}_X(\alpha)),$$
similarly to the inductive case for terms in Definition 2.6. Equivalently,
$$\mathcal{M} \models_\alpha R(t_1, \ldots, t_n) \ :\Longleftrightarrow\ ((t_1)^{\mathcal{M}}_X(\alpha), \ldots, (t_n)^{\mathcal{M}}_X(\alpha)) \in R^{\mathcal{M}},$$
$$R(t_1, \ldots, t_n)^{\mathcal{M}}_X := ((t_1)^{\mathcal{M}}_X, \ldots, (t_n)^{\mathcal{M}}_X)^{-1}(R^{\mathcal{M}})$$
(where the RHS on the last line denotes the preimage of $R^{\mathcal{M}} \subseteq M^n$ under the function $((t_1)^{\mathcal{M}}_X, \ldots, (t_n)^{\mathcal{M}}_X) : M^X \to M^n$ whose coordinates are the $(t_i)^{\mathcal{M}}_X : M^X \to M$).

  When R is the equality symbol $=$, we always take $=^{\mathcal{M}}$ to be the equality relation, i.e., the set or function $=_M$ defined in the discussion before Convention 2.1, or equivalently,
$$\mathcal{M} \models_\alpha s = t \ :\Longleftrightarrow\ s^{\mathcal{M}}_X(\alpha) = t^{\mathcal{M}}_X(\alpha).$$

- The connective cases are the same as in propositional logic: for $\phi, \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$,
$$(\phi \wedge \psi)^{\mathcal{M}}_X(\alpha) := \min(\phi^{\mathcal{M}}_X(\alpha), \psi^{\mathcal{M}}_X(\alpha)),$$
$$(\phi \vee \psi)^{\mathcal{M}}_X(\alpha) := \max(\phi^{\mathcal{M}}_X(\alpha), \psi^{\mathcal{M}}_X(\alpha)),$$
$$(\neg\phi)^{\mathcal{M}}_X(\alpha) := 1 - \phi^{\mathcal{M}}_X(\alpha),$$
$$\top^{\mathcal{M}}_X(\alpha) := 1,$$
$$\bot^{\mathcal{M}}_X(\alpha) := 0.$$
Equivalently,
$$\mathcal{M} \models_\alpha \phi \wedge \psi \ :\Longleftrightarrow\ \mathcal{M} \models_\alpha \phi \text{ and } \mathcal{M} \models_\alpha \psi, \qquad (\phi \wedge \psi)^{\mathcal{M}}_X := \phi^{\mathcal{M}}_X \cap \psi^{\mathcal{M}}_X,$$
$$\mathcal{M} \models_\alpha \phi \vee \psi \ :\Longleftrightarrow\ \mathcal{M} \models_\alpha \phi \text{ or } \mathcal{M} \models_\alpha \psi, \qquad (\phi \vee \psi)^{\mathcal{M}}_X := \phi^{\mathcal{M}}_X \cup \psi^{\mathcal{M}}_X,$$
$$\mathcal{M} \models_\alpha \neg\phi \ :\Longleftrightarrow\ \mathcal{M} \not\models_\alpha \phi, \qquad (\neg\phi)^{\mathcal{M}}_X := M^X \setminus \phi^{\mathcal{M}}_X,$$
$$\mathcal{M} \models_\alpha \top \quad \text{always}, \qquad \top^{\mathcal{M}}_X := M^X,$$
$$\mathcal{M} \models_\alpha \bot \quad \text{never}, \qquad \bot^{\mathcal{M}}_X := \varnothing.$$

- Finally, suppose $\exists x\,\phi \in \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$; then from the definition of free variables, we have $X \supseteq \mathrm{FV}(\exists x\,\phi) = \mathrm{FV}(\phi) \setminus \{x\}$, whence $X \cup \{x\} \supseteq \mathrm{FV}(\phi)$, so that (by the IH) we may assume given the interpretation of ϕ under any $(X \cup \{x\})$-variable assignment. We then define

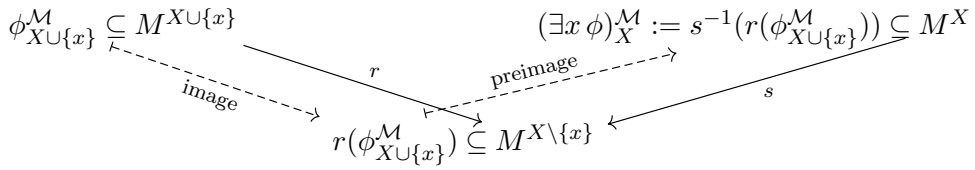
$$(\exists x\,\phi)^{\mathcal{M}}_X(\alpha) := \max_{a \in M} \phi^{\mathcal{M}}_{X \cup \{x\}}(\alpha\langle x \mapsto a\rangle)$$

(where by convention, the max is 0 if $M = \varnothing$). In other words, we interpret $\exists x\,\phi$ as true iff there is some a we can assign to x (ignoring any previous assignment in α) to make ϕ true:

$$\mathcal{M} \models_\alpha \exists x\,\phi \iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle x \mapsto a\rangle} \phi.$$

(Note that the $\exists$ on the RHS is a "meta" $\exists$, whereas the one on the LHS is a symbol! For clarity, we could have written it out in words as "there exists", similar to our use of "and", "or" above; but we find this common abbreviation too convenient to resist, and so will rely on you to recognize the distinction based on context.)

(Aside) The definition of $(\exists x\,\phi)^{\mathcal{M}}_X \subseteq M^X$ as a set is a bit more involved to describe. Start with $\phi^{\mathcal{M}}_{X\cup\{x\}} \subseteq M^{X\cup\{x\}}$, and consider the restriction function $r : M^{X\cup\{x\}} \to M^{X\setminus\{x\}}$ which forgets about the value of a variable assignment at x; then the image $r(\phi^{\mathcal{M}}_{X\cup\{x\}}) \subseteq M^{X\setminus\{x\}}$ is the set of assignments which can be extended with an assignment to x satisfying ϕ. But since we also need to ignore any previous assignment to x, consider also the restriction function $s : M^X \to M^{X\setminus\{x\}}$ (which is either the identity function if $x \notin X$, or otherwise is the same as r); then the preimage $s^{-1}(r(\phi^{\mathcal{M}}_{X\cup\{x\}})) \subseteq M^X$ is the set of assignments for which we can first discard any previous assignment to x (i.e., apply s), and *then* extend with a new assignment to x satisfying ϕ, which exactly yields $(\exists x\,\phi)^{\mathcal{M}}_X$. The following diagram depicts this two-step construction of $(\exists x\,\phi)^{\mathcal{M}}_X$ from $\phi^{\mathcal{M}}_{X\cup\{x\}}$:



- Let us also record the interpretation of $\forall x\,\phi := \neg\exists x\,\neg\phi$, derived from that of $\exists, \neg$:

$$(\forall x\,\phi)^{\mathcal{M}}_X(\alpha) := \min_{a \in M} \phi^{\mathcal{M}}_{X\cup\{x\}}(\alpha\langle x \mapsto a\rangle),$$
$$\mathcal{M} \models_\alpha \forall x\,\phi \iff \forall a \in M,\ \mathcal{M} \models_{\alpha\langle x\mapsto a\rangle} \phi.$$

(The definition of $(\forall x\,\phi)^{\mathcal{M}}_X \subseteq M^X$ as a set is similar to that of $(\exists x\,\phi)^{\mathcal{M}}_X$, involving a "coimage" rather than an image, and is left to you as an Exercise.)

Example 2.13. In the $\mathcal{A}_{\mathsf{ordfield}}$-structure $\mathcal{R} := \mathbb{R}$ with the usual interpretations, consider the sentence $\forall x\,\exists y\,(\neg(y = x) \wedge (x \leq y))$ under the empty variable assignment:

$$\begin{aligned}
&\mathcal{R} \models_\varnothing \forall x\,\exists y\,(\neg(y = x) \wedge (x \leq y)) \\
&\iff \forall a \in \mathbb{R},\ \mathcal{R} \models_{x\mapsto a} \exists y\,(\neg(y = x) \wedge (x \leq y)) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } \mathcal{R} \models_{x\mapsto a, y\mapsto b} \neg(y = x) \wedge (x \leq y) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } (\mathcal{R} \not\models_{x\mapsto a, y\mapsto b} y = x \text{ and } \mathcal{R} \models_{x\mapsto a, y\mapsto b} x \leq y) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } (b \neq a \text{ and } a \leq b) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } a < b
\end{aligned}$$

which is clearly true, since given a, we can take $b := 1$.

We can also regard $\forall x\,\exists y\,(\neg(y = x) \wedge x \leq y)$ as having free variables from $\{x, y\}$, hence interpret it under some assignment of those variables, e.g.,

$$\begin{aligned}
&\mathcal{R} \models_{x\mapsto 3, y\mapsto 2} \forall x\,\exists y\,(\neg(y = x) \wedge (x \leq y)) \\
&\iff \forall a \in \mathbb{R},\ \mathcal{R} \models_{x\mapsto a, y\mapsto 2} \exists y\,(\neg(y = x) \wedge (x \leq y)) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } \mathcal{R} \models_{x\mapsto a, y\mapsto b} \neg(y = x) \wedge (x \leq y) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } (\mathcal{R} \not\models_{x\mapsto a, y\mapsto b} y = x \text{ and } \mathcal{R} \models_{x\mapsto a, y\mapsto b} x \leq y) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } (b \neq a \text{ and } a \leq b) \\
&\iff \forall a \in \mathbb{R},\ \exists b \in \mathbb{R} \text{ s.t. } a < b
\end{aligned}$$

which is true as before. Note that the original values of x, y are "overridden": e.g., in the second line, we used

$$(x \mapsto 3, y \mapsto 2)\langle x \mapsto a\rangle = (x \mapsto a, y \mapsto 2).$$

In particular, the formula $x \leq y$ is never evaluated with the original values $x = 3$, $y = 2$.

Exercise 2.14. Determine whether or not $\mathcal{R} \models_{x\mapsto 1, y\mapsto 2} \forall x\,((\exists x\,(y \cdot y = x)) \to (\exists y\,(y \cdot y = x)))$.

Exercise 2.15. Let ϕ be any formula with free variables from X, and let $x \in X$. Verify that for any $\mathcal{M}$ and $\alpha : X \to M$,

$$\mathcal{M} \models_\alpha (\forall x\,\phi) \to \phi.$$

(This reflects the common situation where you find yourself knowing that "for all x, ...", and you use this to conclude that ... holds for an already fixed x. We will see the inference rule that formalizes this way of reasoning in Example 4.27 below.) What about

$$\mathcal{M} \models_\alpha (\exists x\,\phi) \to \phi,$$
$$\mathcal{M} \models_\alpha \phi \to (\forall x\,\phi),$$
$$\mathcal{M} \models_\alpha \phi \to (\exists x\,\phi)?$$

For a sentence ϕ, we write $\mathcal{M} \models \phi$ instead of $\mathcal{M} \models_\varnothing \phi$. We call a sentence ϕ a **semantic tautology**, written

$$\models \phi,$$

if it is satisfied by all $\mathcal{A}$-structures. We say that ϕ is **satisfiable** if it is satisfied by *some* structure, and **unsatisfiable** otherwise.

For a general ϕ with free variables from X, we may say that ϕ is a **semantic tautology**, written

$$\models_X \phi,$$

if it satisfied by all $\mathcal{A}$-structures under all variable assignments to X; for $X = \{x_1, \ldots, x_n\}$ finite, this is the same as saying that the sentence $\forall x_1 \cdots \forall x_n\,\phi$ is a semantic tautology. We also say that ϕ **semantically implies** ψ if $\phi \to \psi$ is a semantic tautology.

Example 2.16. $\forall x\,\forall y\,\forall z\,((x = y) \wedge (y = z) \to (x = z))$ is a semantic tautology, since for any $\mathcal{M}$,

$$\mathcal{M} \models \forall x\,\forall y\,\forall z\,((x = y) \wedge (y = z) \to (x = z)) \iff \forall a, b, c \in M\,(a = b \text{ and } b = c \implies a = c)$$

which is true by transitivity of equality. (So $(x = y) \wedge (y = z)$ semantically implies $x = z$.)

Example 2.17. $(x + y) + z = x + (y + z)$ is *not* a semantic tautology (for any $\mathcal{A}$ containing a binary function symbol $+$, e.g., $\mathcal{A}_{\mathsf{ordfield}}$), since its interpretation in $\mathcal{M} := \mathbb{R}$ with $+^{\mathcal{M}} :=$ subtraction is false, under the assignment $x \mapsto 0$, $y \mapsto 1$, and $z \mapsto 1$, say.

Remark 2.18. Unlike in propositional logic, it is almost never possible to determine semantic truth by drawing a truth table (since there are infinitely many $\mathcal{A}$-structures, even if $\mathcal{A}$ is finite).

Remark 2.19. There is a very subtle ambiguity in an edge case in the definition of "semantic tautology" if we allow free variables: if ϕ has free variables from $X \subseteq Y$, then does whether ϕ is a semantic tautology depend on whether we regard it has having free variables from X or from Y? In other words, is $\models_X \phi \iff \models_Y \phi$? Note that this issue is not quite covered by Proposition 2.10, which only says that the truth value of ϕ in a *particular* structure $\mathcal{M}$ under a *particular* variable assignment α only depends on the free variables actually occurring in ϕ; whereas the definition of "semantic tautology" involves a "for all" quantifier over $\mathcal{M}, \alpha$.

Indeed, consider the sentence $\phi := \exists x \top$ (mentioned in Example 1.10). For any $\mathcal{M}$, we have

$$\begin{aligned} \mathcal{M} \models \phi &\iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{x \mapsto a} \top \\ &\iff \exists a \in M, \end{aligned}$$

i.e., ϕ asserts that the underlying set is nonempty. So ϕ is not a tautology *as a sentence*, i.e., $\not\models_\varnothing \phi$, since ϕ is false in an empty structure. However, if we instead regard ϕ as having free variables from some *nonempty* X, then ϕ *is* a tautology, i.e., $\models_X \phi$, since for any $\mathcal{M}$ and variable assignment $\alpha \in M^X$, since X is nonempty, M must be nonempty, whence $\mathcal{M} \models \phi$.

Because of this issue, we should really say "ϕ is a tautology over X", or just write the unambiguous notation $\models_X \phi$. However, the following shows that this issue really does only come up in edge cases:

Exercise 2.20.

(a) Show that if $X \subseteq Y$, then $\models_X \phi \implies \models_Y \phi$.
(b) Show that the converse holds assuming either every empty structure satisfies ϕ, or $X \neq \varnothing$.

2.3. **Theories.** Let $\mathcal{A}$ be a first-order signature. An **$\mathcal{A}$-theory** $\mathcal{T}$ is a set of $\mathcal{A}$-*sentences*, which are called **axioms** of $\mathcal{T}$. An $\mathcal{A}$-structure $\mathcal{M}$ is a **model** of $\mathcal{T}$ if it satisfies every axiom in $\mathcal{T}$, written

$$\mathcal{M} \models \mathcal{T} :\iff \forall \phi \in \mathcal{T}\, (\mathcal{M} \models \phi).$$

Let

$$\mathrm{Mod}(\mathcal{T}) := \{\mathcal{A}\text{-structures } \mathcal{M} \mid \mathcal{M} \models \mathcal{T}\}$$

denote the collection[3] of all models of $\mathcal{T}$; we say that $\mathcal{T}$ **axiomatizes** $\mathrm{Mod}(\mathcal{T})$.

Example 2.21. The **theory of (simple undirected) graphs** is the $\mathcal{A}_{\mathsf{graph}} = \{E\}$-theory

$$\mathcal{T}_{\mathsf{graph}} := \{\forall x \neg E(x,x),\ \forall x \forall y\, (E(x,y) \to E(y,x))\}.$$

An $\mathcal{A}_{\mathsf{graph}}$-structure $\mathcal{M} = (M, E^{\mathcal{M}})$, where $E^{\mathcal{M}} \subseteq M^2$, is a model of $\mathcal{T}_{\mathsf{graph}}$ iff

$$\forall a \in M\, ((a,a) \notin E^{\mathcal{M}}), \qquad \forall (a,b) \in E^{\mathcal{M}}\, ((b,a) \in E^{\mathcal{M}}).$$

Example 2.22. The **theory of posets (partially ordered sets)** is the $\mathcal{A}_{\mathsf{poset}} = \{\leq\}$-theory

$$\begin{aligned} \mathcal{T}_{\mathsf{poset}} := \{&\forall x\, (x \leq x), \\ &\forall x \forall y \forall z\, ((x \leq y) \wedge (y \leq z) \to (x \leq z)), \\ &\forall x \forall y\, ((x \leq y) \wedge (y \leq x) \to (x = y))\}. \end{aligned}$$

An $\mathcal{A}_{\mathsf{poset}}$-structure $\mathcal{M} = (M, \leq^{\mathcal{M}})$ is a model of $\mathcal{T}_{\mathsf{poset}}$ iff $\leq^{\mathcal{M}}$ is a reflexive, transitive, and antisymmetric (meaning $a \leq^{\mathcal{M}} b \leq^{\mathcal{M}} a \implies a = b$) binary relation on M.

Example 2.23. The **theory of totally ordered sets** is the $\mathcal{A}_{\mathsf{poset}}$-theory

$$\mathcal{T}_{\mathsf{toset}} := \mathcal{T}_{\mathsf{poset}} \cup \{\forall x \forall y\, ((x \leq y) \vee (y \leq x))\}.$$

[3]Unlike in propositional logic, $\mathrm{Mod}(\mathcal{T})$ is generally a proper class, not a set.

Example 2.24. The **theory of equivalence relations** is the $\mathcal{A}_{\mathsf{equiv}} = \{\sim\}$-theory consisting of the first two axioms of $\mathcal{T}_{\mathsf{poset}}$ together with the last axiom ("symmetry") of $\mathcal{T}_{\mathsf{graph}}$, with the relation symbols replaced by $\sim$.

As these examples show, it is often useful to "modularize" theories into groups of related axioms.

Example 2.25. The **theory of abelian groups** is the theory over the signature $\mathcal{A}_{\mathsf{abgrp}} = \{+, 0, -\}$, consisting of $(2, 0, 1)$-ary function symbols respectively, given by

$$\begin{aligned}\mathcal{T}_{\mathsf{abgrp}} := \{&\forall x\,\forall y\,\forall z\,((x+y)+z = x+(y+z)) && (+\text{-associativity}),\\ &\forall x\,(x+0=x) && (+\text{-identity}),\\ &\forall x\,\forall y\,(x+y=y+x) && (+\text{-commutativity}),\\ &\forall x\,(x+(-x)=0)\} && (+\text{-inverse}).\end{aligned}$$

Models include $\mathbb{Z}$, $2\mathbb{Z} = \{2n \mid n \in \mathbb{Z}\}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{R}^2$, $\mathbb{R}^3$, ... with the usual operations.

The **theory of commutative rings** is the $\mathcal{A}_{\mathsf{ring}} = \mathcal{A}_{\mathsf{field}} = \{+, 0, -, \cdot, 1\}$-theory

$$\begin{aligned}\mathcal{T}_{\mathsf{commring}} := \mathcal{T}_{\mathsf{abgrp}} \cup \{&\forall x\,\forall y\,\forall z\,((x\cdot y)\cdot z = x\cdot(y\cdot z)) && (\cdot\text{-associativity}),\\ &\forall x\,(x\cdot 1 = x) && (\cdot\text{-identity}),\\ &\forall x\,\forall y\,(x\cdot y = y\cdot x) && (\cdot\text{-commutativity}),\\ &\forall x\,\forall y\,\forall z\,(x\cdot(y+z) = x\cdot y + x\cdot z)\} && (\text{distributivity}).\end{aligned}$$

Models include $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}, \mathcal{C}(\mathbb{R}) := \{\text{continuous functions } \mathbb{R} \to \mathbb{R}\}$ (with pointwise operations).

The **theory of fields** is the $\mathcal{A}_{\mathsf{field}}$-theory

$$\begin{aligned}\mathcal{T}_{\mathsf{field}} := \mathcal{T}_{\mathsf{commring}} \cup \{&\neg(0=1),\\ &\forall x\,(\neg(x=0) \to \exists y\,(x\cdot y = 1))\}.\end{aligned}$$

Models include $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ (not $\mathcal{C}(\mathbb{R})$, since e.g., $\sin \neq 0$ but does not have a multiplicative inverse).

The **theory of ordered fields** is the $\mathcal{A}_{\mathsf{ordfield}}$-theory

$$\begin{aligned}\mathcal{T}_{\mathsf{ordfield}} := \mathcal{T}_{\mathsf{field}} \cup \mathcal{T}_{\mathsf{toset}} \cup \{&\forall x\,\forall y\,\forall z\,((x\le y) \to (x+z \le y+z)),\\ &\forall x\,\forall y\,\forall z\,((x\le y)\wedge(0\le z) \to (x\cdot z \le y\cdot z))\}.\end{aligned}$$

Models include $\mathbb{Q}, \mathbb{R}$ (not $\mathbb{C}$).

Example 2.26. The **theory of ($\mathbb{R}$-)vector spaces** is the *uncountable* $\mathcal{A}_{\mathsf{vec}}$-theory

$$\mathcal{T}_{\mathsf{vec}} := \mathcal{T}_{\mathsf{abgrp}} \cup \left\{ \begin{array}{l} \forall x\,((ab)\cdot x = a\cdot(b\cdot x)) \\ \forall x\,(1\cdot x = x) \\ \forall x\,((a+b)\cdot x = a\cdot x + b\cdot x) \\ \forall x\,\forall y\,(a\cdot(x+y) = a\cdot x + a\cdot y) \end{array} \middle| \; a, b \in \mathbb{R} \right\}.$$

Note that $\forall x$ quantifies over elements of the structure (i.e., vectors), while to express laws which hold for each scalar, we need infinite families of axioms, one for each scalar. (Note also the differing roles of a, b, x in e.g., the third axiom, in which the terms on either side have the tree structures

In particular, note that $+$ in $a + b$ does *not* refer to the symbol $+ \in \mathcal{A}_{\mathsf{vec}}$ (which only adds vectors), but rather to the usual addition in $\mathbb{R}$.)

Example 2.27. Here is another example of the same kind of idea. Suppose we wish to axiomatize structures which are sets M equipped with an injective sequence $f : \mathbb{N} \to M$ of distinct elements. Since the domain $\mathbb{N}$ does not consist of elements of M, we should not use a unary function symbol f; rather, for each $n \in \mathbb{N}$, we treat $f(n)$ as a *single constant symbol*, and put

$$\mathcal{A} := \{f(0), f(1), f(2), \dots\}.$$

Thus an $\mathcal{A}$-structure $\mathcal{M}$ consists of a set M together with elements $f(0)^{\mathcal{M}}, f(1)^{\mathcal{M}}, \dots \in M$. Now to enforce injectivity, we use the theory

$$\mathcal{T} := \{\neg(f(m) = f(n)) \mid m \neq n \in \mathbb{N}\}.$$

(Note that it is not possible to enforce *surjectivity* of f via a first-order theory. Intuitively, this is because we would need to say $\forall x\,((x = f(0)) \vee (x = f(1)) \vee \cdots)$. We can prove that surjectivity is not axiomatizable, once we have the compactness theorem for first-order logic; see Theorem 5.18.)

Here are some degenerate examples of theories:

Example 2.28. The $\varnothing$-theory $\varnothing$ axiomatizes the class of all sets (i.e., $\varnothing$-structures).

Example 2.29. The $\varnothing$-theory $\mathcal{T} = \{\forall x\,\forall y\,(x = y)\}$ axiomatizes the class of all sets M with $|M| \leq 1$.

Example 2.30. The $\{P\}$-theory $\varnothing$, where P is a 0-ary relation symbol, axiomatizes the class of all pairs $(M, P^{\mathcal{M}})$ where M is a set and $P^{\mathcal{M}} \in \{0, 1\}$.

We say that a sentence ϕ is a **semantic consequence** of $\mathcal{T}$, or **semantically implied** by $\mathcal{T}$, if it holds in every model of $\mathcal{T}$, written

$$\mathcal{T} \models \phi \;:\Longleftrightarrow\; \forall \mathcal{M} \models \mathcal{T}\,(\mathcal{M} \models \phi).$$

As before (keeping in mind Remark 2.19), we may extend this to formulas with free variables from X, denoted $\mathcal{T} \models_X \phi$, by also considering all variable assignments $\alpha : X \to M$. We also say that $\mathcal{T}$ is **satisfiable** if it has a model, and **unsatisfiable** otherwise.

Example 2.31. We have

$$\mathcal{T}_{\mathsf{abgrp}} \models \forall x\,\forall y\,\forall z\,((x + z = y + z) \to (x = y)),$$

since given an abelian group $\mathcal{M} \models \mathcal{T}_{\mathsf{abgrp}}$, the interpretation of this sentence in $\mathcal{M}$ says that

$$\forall a, b, c \in M\,(a +^{\mathcal{M}} c = b +^{\mathcal{M}} c \implies a = b).$$

To prove this "externally", let $a, b, c \in M$ such that

$$a +^{\mathcal{M}} c = b +^{\mathcal{M}} c.$$

Adding $-^{\mathcal{M}} c$ to both sides yields

$$(a +^{\mathcal{M}} c) +^{\mathcal{M}} (-^{\mathcal{M}} c) = (b +^{\mathcal{M}} c) +^{\mathcal{M}} (-^{\mathcal{M}} c).$$

By the interpretation of the associativity axiom of $\mathcal{T}_{\mathsf{abgrp}}$ in $\mathcal{M}$,

$$a +^{\mathcal{M}} (c +^{\mathcal{M}} (-^{\mathcal{M}} c)) = b +^{\mathcal{M}} (c +^{\mathcal{M}} (-^{\mathcal{M}} c)).$$

By the inverse axiom,

$$a +^{\mathcal{M}} 0^{\mathcal{M}} = b +^{\mathcal{M}} 0^{\mathcal{M}}.$$

By the identity axiom,

$$a = b.$$

Later, we will formalize this kind of reasoning in the proof system for first-order logic (see Example 4.28 and HW10).

2.4. **Homomorphisms and isomorphisms.** These are an entirely new feature of first-order logic (compared to propositional logic). In propositional logic, to say that two models $m, n : \mathcal{A} \to \{0,1\}$ are "the same" just means that they are equal as functions. But in first-order logic, it is possible for two models $\mathcal{M}, \mathcal{N}$ to look "the same" for all intents and purposes, without actually being equal:

Example 2.32. You've probably seen the formal construction of the field of rationals $\mathbb{Q}$ from the integers $\mathbb{Z}$: we essentially define a rational $q \in \mathbb{Q}$ to mean a fraction a/b (which is formally just an ordered pair (a, b)) where $a, b \in \mathbb{Z}$, $b \neq 0$. But since it is possible for two different fractions to represent the same rational, we need to quotient by an equivalence relation $(a, b) \sim (c, d) \iff ad = bc$. Alternatively, we could choose to always use the "reduced form" where a, b are coprime and $b \geq 1$. This yields two different fields, call them $\mathbb{Q}_1$ and $\mathbb{Q}_2$, which are not *equal*, since e.g., the constant $1^{\mathbb{Q}_1} = \{(1,1), (2,2), (-3,-3), \dots\}$ is an equivalence class of pairs of integers whereas $1^{\mathbb{Q}_2} = (1,1)$ is a single pair of integers; but clearly $\mathbb{Q}_1, \mathbb{Q}_2$ should be interchangeable as fields.

This "interchangeability as structures" is formalized as the notion of *isomorphism*, a structure-preserving 1–1 correspondence (i.e., bijection). First, we consider the more general one-way notion of a structure-preserving function or *homomorphism*.

Definition 2.33. Let $\mathcal{A}$ be a signature, $\mathcal{M}, \mathcal{N}$ be two $\mathcal{A}$-structures. An **$\mathcal{A}$-homomorphism** $h : \mathcal{M} \to \mathcal{N}$ is a function $h : M \to N$ between their underlying sets which preserves the structure:

- for each n-ary function symbol $f \in \mathcal{A}^n_{\text{fun}}$, we have
$$h(f^{\mathcal{M}}(a_1, \dots, a_n)) = f^{\mathcal{N}}(h(a_1), \dots, h(a_n))$$
for all $a_1, \dots, a_n \in M$ (we say that h **preserves (the interpretation of)** f);
- for each n-ary relation symbol $R \in \mathcal{A}^n_{\text{rel}}$, we have
$$(a_1, \dots, a_n) \in R^{\mathcal{M}} \implies (h(a_1), \dots, h(a_n)) \in R^{\mathcal{N}},$$
or equivalently, treating relations as characteristic functions,
$$R^{\mathcal{M}}(a_1, \dots, a_n) \leq R^{\mathcal{N}}(h(a_1), \dots, h(a_n)).$$
Note the $\leq$, not $=$! (We say that h **preserves (the interpretation of)** R.)

Example 2.34. For two $\mathcal{A}_{\mathsf{abgrp}}$-structures $\mathcal{M}, \mathcal{N}$, a homomorphism $h : \mathcal{M} \to \mathcal{N}$ has to obey

$$(*) \qquad \begin{aligned} h(a +^{\mathcal{M}} b) &= h(a) +^{\mathcal{N}} h(b), \\ h(0^{\mathcal{M}}) &= 0^{\mathcal{N}}, \\ h(-^{\mathcal{M}} a) &= -^{\mathcal{N}} h(a). \end{aligned}$$

Note that these conditions have nothing to do with what axioms $\mathcal{M}, \mathcal{N}$ satisfy! If $\mathcal{M}, \mathcal{N}$ happen to be abelian groups (i.e., models of $\mathcal{T}_{\mathsf{abgrp}}$), then we call h an **abelian group homomorphism**.

Example 2.35. For two vector spaces $\mathcal{M}, \mathcal{N} \models \mathcal{T}_{\mathsf{vec}}$, a homomorphism $h : \mathcal{M} \to \mathcal{N}$ has to obey the above, as well as, for each $r \in \mathbb{R}$ and $a \in M$,

$$(\dagger) \qquad h(r \cdot^{\mathcal{M}} a) = r \cdot^{\mathcal{N}} h(a).$$

Of course, homomorphisms of vector spaces are usually called **linear transformations**.

Linear transformations are usually defined (in a linear algebra class, say) by requiring only the two conditions $(*)$ and $(\dagger)$. This is an accident specific to vector spaces (and certain other structures), where preservation of certain structure implies preservation of others; the general notion of homomorphism, which works for all structures, requires preservation of all parts of the structure. In particular, note that the following proof doesn't work for arbitrary $\mathcal{A}_{\mathsf{abgrp}}$-structures; it uses the axioms in $\mathcal{T}_{\mathsf{abgrp}}$ in a key way:

Proposition 2.36. If a function $h : \mathcal{M} \to \mathcal{N}$ between abelian groups preserves $+$, then it also preserves $0, -$, i.e., is an abelian group homomorphism.

(Thus, if $\mathcal{M}, \mathcal{N}$ are vector spaces and h also preserves each $r\cdot$, then h is a linear transformation.)

Proof. To show $h(0^{\mathcal{M}}) = 0^{\mathcal{N}}$: from preservation of $+$, we have

$$h(0^{\mathcal{M}} +^{\mathcal{M}} 0^{\mathcal{M}}) = h(0^{\mathcal{M}}) +^{\mathcal{N}} h(0^{\mathcal{M}}).$$

By the identity axiom for $+$ in $\mathcal{T}_{\mathsf{abgrp}}$ applied to LHS (twice, in both $\mathcal{M}, \mathcal{N}$), we get

$$h(0^{\mathcal{M}}) + 0^{\mathcal{N}} = h(0^{\mathcal{M}}) +^{\mathcal{N}} h(0^{\mathcal{M}}).$$

Now applying the commutative law to the LHS and the cancellation law from Example 2.31 yields

$$0^{\mathcal{N}} = h(0^{\mathcal{M}}).$$

Preservation of $-$ is similar, and left as an **Exercise**. □

Exercise 2.37. An **abelian monoid** is an $\mathcal{A}_{\mathsf{abmon}} := \{+, 0\}$-structure which only obeys the first three axioms in $\mathcal{T}_{\mathsf{abgrp}}$ (call this theory $\mathcal{T}_{\mathsf{abmon}}$). Show that a function between abelian monoids which only preserves $+$ need not be an abelian monoid homomorphism.

Example 2.38. For two posets $\mathcal{M}, \mathcal{N} \models \mathcal{T}_{\mathsf{poset}}$, a homomorphism $h : \mathcal{M} \to \mathcal{N}$ has to obey

$$a \leq^{\mathcal{M}} b \implies h(a) \leq^{\mathcal{N}} h(b).$$

This is usually called an **order-preserving** or **monotone** function. For example,

$$\begin{aligned} \exp : \mathbb{R} &\longrightarrow \mathbb{R} \\ x &\longmapsto e^x \end{aligned}$$

is monotone when $\mathbb{R}$ is equipped with the usual $\leq$. For a non-numeric example, for any function $f : X \to Y$ between sets, taking preimage yields a monotone function between their powersets:

$$f^{-1} : \mathcal{P}(Y) \longrightarrow \mathcal{P}(X).$$

(Note that we do *not* require h to be strictly order-preserving, i.e., $a < b \implies h(a) < h(b)$.)

Exercise 2.39. Show that $f^{-1} : \mathcal{P}(Y) \to \mathcal{P}(X)$ is strictly order-preserving iff f is surjective.

Definition 2.40. A homomorphism $h : \mathcal{M} \to \mathcal{N}$ is an **$\mathcal{A}$-isomorphism**, written $h : \mathcal{M} \cong \mathcal{N}$, if

(i) it is a bijection, i.e., it has an inverse $h^{-1} : N \to M$, which *a priori* is just a function;
(ii) h^{-1} is also a homomorphism.

You may have seen examples of structures (e.g., in linear algebra) where condition (ii) is unnecessary, i.e., every bijective homomorphism is automatically an isomorphism. This is the case only for some types of structures, as the following shows:

Example 2.41. Fix $n \in \mathbb{N}$, and let $\{0,1\}^n$, the set of all length n strings of bits, be partially ordered coordinatewise, i.e.,

$$(a_0, \ldots, a_{n-1}) \leq (b_0, \ldots, b_{n-1}) \ :\iff\ a_0 \leq b_0 \text{ and } \cdots \text{ and } a_{n-1} \leq b_{n-1}.$$

Thus $\{0,1\}^n$ becomes a poset. Thinking of each such finite string as the binary digits of a number between $0, 2^n - 1$, we get a bijection

$$\begin{aligned} h : \{0,1\}^n &\longrightarrow \{0, 1, \ldots, 2^n - 1\} \\ (a_0, \ldots, a_{n-1}) &\longmapsto a_0 + 2a_1 + 4a_2 + \cdots + 2^{n-1}a_{n-1}, \end{aligned}$$

which is clearly order-preserving when $\{0, 1, \ldots, 2^n - 1\}$ is equipped with the usual total ordering from $\mathbb{Z}$, but is not an isomorphism of posets, since e.g., when $n = 2$,

$$h^{-1}(1) = (1,0) \not\leq (0,1) = h^{-1}(2).$$

Proposition 2.42. If $\mathcal{A}$ consists only of function symbols, then every bijective $\mathcal{A}$-homomorphism $h : \mathcal{M} \to \mathcal{N}$ is an isomorphism, i.e., h^{-1} is automatically also a homomorphism.

Proof. Let $f \in \mathcal{A}^n_{\text{fun}}$; we must show that for $a_1, \dots, a_n \in N$,

$$h^{-1}(f^{\mathcal{N}}(a_1, \dots, a_n)) = f^{\mathcal{M}}(h^{-1}(a_1), \dots, h^{-1}(a_n)).$$

Since h is a homomorphism, we have

$$\begin{aligned} h(f^{\mathcal{N}}(h^{-1}(a_1), \dots, h^{-1}(a_n))) &= f^{\mathcal{N}}(h(h^{-1}(a_1)), \dots, h(h^{-1}(a_n))) \\ &= f^{\mathcal{N}}(a_1, \dots, a_n). \end{aligned}$$

Applying h^{-1} to both sides yields the result. □

Exercise 2.43. Show that every bijective homomorphism between *totally* ordered sets is an isomorphism.

2.5. **Preservation of formulas.** A homomorphism by definition preserves "atomic terms", i.e., function symbols, as well as atomic formulas, i.e., relation symbols. We now show that they preserve most other derived terms and formulas, while isomorphisms preserve *everything*. It is convenient to split this proof into several lemmas, one for each type of operation used to build terms/formulas:

Lemma 2.44. A homomorphism $h : \mathcal{M} \to \mathcal{N}$ **preserves the interpretation of terms**: for each term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ and variable assignment $\alpha : X \to M$, we have

$$h(t^{\mathcal{M}}(\alpha)) = t^{\mathcal{N}}(h \circ \alpha).$$

(If we think of $\alpha \in M^X$ as an "X-ary tuple" $(\alpha(x))_{x \in X}$, then $h \circ \alpha = (h(\alpha(x)))_{x \in X}$ is the result of applying h to each coordinate, analogously to the preservation of functions in Definition 2.33.)

Proof. By induction on t.

- For a variable $t = x \in X$, we have

$$\begin{aligned} h(x^{\mathcal{M}}(\alpha)) &= h(\alpha(x)) && \text{by definition of } x^{\mathcal{M}} \text{ (see Definition 2.6)} \\ &= x^{\mathcal{N}}(h \circ \alpha) && \text{by definition of } x^{\mathcal{N}}. \end{aligned}$$

- For $t = f(t_1, \dots, t_n)$ where $f \in \mathcal{A}^n_{\text{fun}}$ and $t_1, \dots, t_n \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we have

$$\begin{aligned} h(f(t_1, \dots, t_n)^{\mathcal{M}}(\alpha)) &= h(f^{\mathcal{M}}(t_1^{\mathcal{M}}(\alpha), \dots, t_n^{\mathcal{M}}(\alpha))) && \text{by definition of } f(t_1, \dots, t_n)^{\mathcal{M}} \\ &= f^{\mathcal{N}}(h(t_1^{\mathcal{M}}(\alpha)), \dots, h(t_n^{\mathcal{M}}(\alpha))) && \text{since } h \text{ is a homomorphism} \\ &= f^{\mathcal{N}}(t_1^{\mathcal{N}}(h \circ \alpha), \dots, t_n^{\mathcal{N}}(h \circ \alpha)) && \text{by IH} \\ &= f(t_1, \dots, t_n)^{\mathcal{N}}(h \circ \alpha) && \text{by definition of } f(t_1, \dots, t_n)^{\mathcal{N}}. \end{aligned}$$

□

Example 2.45. For a homomorphism $h : \mathcal{M} \to \mathcal{N}$ between vector spaces, letting, say,

$$t := 3 \cdot x + 4 \cdot (5 \cdot y + (-z)),$$

the above lemma says that for all $a, b, c \in M$, (omitting superscripts $^{\mathcal{M}}$, $^{\mathcal{N}}$ for clarity)

$$h(3 \cdot a + 4 \cdot (5 \cdot b + (-c))) = 3 \cdot h(a) + 4 \cdot (5 \cdot h(b) + (-h(c))).$$

In other words, we recover the familiar fact that linear transformations preserve linear combinations.

Example 2.46. For a homomorphism $h : \mathcal{M} \to \mathcal{N}$ between commutative rings (see Example 2.25), we can likewise think of a term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A}_{\mathsf{commring}})$ as a polynomial with integer coefficients, e.g.,

$$t = x \cdot x + x \cdot x + x \cdot x + (-x) + 1 + 1$$

would be how we formally write $3x^2 - x + 2$; h then has to preserve evaluation of all such polynomials.

For a formula $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, we likewise say that $h : \mathcal{M} \to \mathcal{N}$ **preserves the interpretation of** ϕ if for each variable assignment $\alpha : X \to M$, we have any of the following equivalent conditions:

$$\alpha \in \phi^{\mathcal{M}}_X \implies h \circ \alpha \in \phi^{\mathcal{N}}_X,$$
$$\phi^{\mathcal{M}}(\alpha) \le \phi^{\mathcal{N}}(h \circ \alpha),$$
$$\mathcal{M} \models_\alpha \phi \implies \mathcal{N} \models_{h\circ\alpha} \phi$$

(compare again with the preservation of relations condition in Definition 2.33).

Lemma 2.47. Homomorphisms preserve the interpretation of atomic formulas.

Proof. As usual, this is similar to the inductive case in the proof for terms, Lemma 2.44. Alternatively, using the $\models$ notation, we have

$$\begin{aligned}
\mathcal{M} \models_\alpha R(t_1, \dots, t_n) &\iff (t_1^{\mathcal{M}}(\alpha), \dots, t_n^{\mathcal{M}}(\alpha)) \in R^{\mathcal{M}} \\
(*) \qquad &\implies (h(t_1^{\mathcal{M}}(\alpha)), \dots, h(t_n^{\mathcal{M}}(\alpha))) \in R^{\mathcal{N}} \quad \text{since } h \text{ is a homomorphism} \\
&\iff (t_1^{\mathcal{N}}(h \circ \alpha), \dots, t_n^{\mathcal{N}}(h \circ \alpha)) \in R^{\mathcal{N}} \quad \text{by Lemma 2.44} \\
&\iff \mathcal{N} \models_{h\circ\alpha} R(t_1, \dots, t_n).
\end{aligned}$$

(When R is the equal sign $=$, step $(*)$ is instead because all functions preserve equality.) □

Lemma 2.48. Let $h : M \to N$ be an arbitrary function.

(a) If h preserves the interpretations of $\phi, \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, then it also preserves $\phi \wedge \psi, \phi \vee \psi$.
(b) h always preserves the interpretations of $\top, \bot$.
(c) If h preserves the interpretation of $\phi \in \mathcal{L}^{X\cup\{x\}}_{\text{form}}(\mathcal{A})$, then it also preserves $\exists x\, \phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$.

Proof. (a) The key point here is that $\min, \max$ are monotone functions: for $\wedge$,

$$\begin{aligned}
(\phi \wedge \psi)^{\mathcal{M}}(\alpha) &= \min(\phi^{\mathcal{M}}(\alpha), \psi^{\mathcal{M}}(\alpha)) \\
&\le \min(\phi^{\mathcal{N}}(h \circ \alpha), \psi^{\mathcal{N}}(h \circ \alpha)) \quad \text{by assumption and monotonicity of min} \\
&= (\phi \wedge \psi)^{\mathcal{N}}(h \circ \alpha);
\end{aligned}$$

similarly for $\vee$. (So this doesn't work for $\neg$, since $x \mapsto 1 - x$ is not monotone.)

(b) is trivial, since $\top, \bot$ always have the same truth value.

(c) We have

$$\begin{aligned}
\mathcal{M} \models_\alpha \exists x\, \phi &\iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle x\mapsto a\rangle} \phi, \\
(*) \qquad \mathcal{N} \models_{h\circ\alpha} \exists x\, \phi &\iff \exists b \in N \text{ s.t. } \mathcal{N} \models_{(h\circ\alpha)\langle x\mapsto b\rangle} \phi;
\end{aligned}$$

and we must show that the former implies the latter. Given the former, since h preserves ϕ, we get

$$\mathcal{N} \models_{h\circ\alpha\langle x\mapsto a\rangle} \phi.$$

Now $h \circ \alpha\langle x \mapsto a\rangle = (h \circ \alpha)\langle x \mapsto h(a)\rangle : X \cup \{x\} \to N$, since both functions map $x \mapsto h(a)$ and all other $y \in X \setminus \{x\}$ to $h(\alpha(y))$. Thus the RHS of $(*)$ holds with $b := h(a)$. □

Proposition 2.49. Homomorphisms preserve the interpretation of **positive-existential** formulas, i.e., formulas built using $\wedge, \vee, \top, \bot, \exists$ (no $\neg$, hence also no $\forall$ or $\to$).

Proof. By induction; the base case is Lemma 2.47, while the inductive cases are Lemma 2.48. □

Example 2.50. Consider the $\mathcal{A}_{\mathsf{commring}}$-formula with one free variable x

$$\phi := \exists y\, (x \cdot y = 1).$$

The interpretation in a commutative ring $\mathcal{M}$ under an assignment $x \mapsto a$ says that a has a multiplicative inverse. Thus, *commutative ring homomorphisms preserve invertibility.*

The following shows that the restriction to positive-existential formulas above is necessary:

Example 2.51. For any subset $N \subseteq M$ of the underlying set of a structure $\mathcal{M}$, closed under $f^{\mathcal{M}}$ for all $f \in \mathcal{A}_{\text{fun}}$, we have a **substructure** $\mathcal{N}$ on N defined by restricting the interpretations in $\mathcal{M}$ of all the symbols in $\mathcal{A}$ (see HW6). The inclusion function $i : N \hookrightarrow M$ is always a homomorphism from such a substructure (with $\iff$ replacing $\implies$ in the preservation of relations in Definition 2.33).

For example, $\mathbb{Z} \subseteq \mathbb{R}$ is a $\{0, 1, \leq\}$-substructure (under the usual interpretations). The sentence

$$\phi := \forall x \left((x \leq 0) \vee (1 \leq x) \right)$$

is obviously true in $\mathbb{Z}$, but not in $\mathbb{R}$, hence is not preserved by the inclusion $i : \mathbb{Z} \hookrightarrow \mathbb{R}$.

Exercise 2.52 (HW6)**.** Show that any *surjective* homomorphism preserves all **positive** formulas, i.e., formulas built using $\wedge, \vee, \top, \bot, \exists, \forall$ (no $\neg$, except those included in $\forall := \neg\exists\neg$).

Exercise 2.53. Find an (easy) example of a surjective homomorphism that fails to preserve $\neg$.

Lemma 2.54. If $h : \mathcal{M} \to \mathcal{N}$ preserves $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ and is a bijection, then h^{-1} preserves $\neg\phi$.

Proof. For $\alpha : X \to N$,

$$\mathcal{N} \models_\alpha \neg\phi \iff \mathcal{N} \not\models_{\alpha = h \circ h^{-1} \circ \alpha} \phi \implies \mathcal{M} \not\models_{h^{-1} \circ \alpha} \phi \iff \mathcal{M} \models_{h^{-1} \circ \alpha} \neg\phi$$

by the contrapositive of the fact that h preserves the interpretation of ϕ. □

Proposition 2.55. Isomorphisms preserve the interpretation of all first-order formulas.

Proof. For an isomorphism $h : \mathcal{M} \to \mathcal{N}$, we know by definition that $h^{-1} : \mathcal{N} \to \mathcal{M}$ is an isomorphism. We show simultaneously that both h, h^{-1} preserve all formulas ϕ by induction on ϕ. The base and non-$\neg$ inductive cases are by Lemma 2.47 and Lemma 2.48, as in Proposition 2.49. In the $\neg$ case, we use Lemma 2.54 and that h^{-1} preserves ϕ to deduce that h preserves $\neg\phi$, and vice-versa. □

We say that $\mathcal{M}, \mathcal{N}$ are **isomorphic** if there is an isomorphism between them, written

$$\mathcal{M} \cong \mathcal{N} \ :\iff \ \exists h : \mathcal{M} \cong \mathcal{N}.$$

Corollary 2.56. If $\mathcal{M} \cong \mathcal{N}$ and $\mathcal{M} \models \mathcal{T}$, then $\mathcal{N} \models \mathcal{T}$.

In other words, every axiomatizable class of structures $\mathcal{K} = \text{Mod}(\mathcal{T})$ has to be closed under $\cong$.

Example 2.57. If $\mathcal{M}$ is a field, and $\mathcal{N}$ is a commutative ring (see Example 2.25) isomorphic to $\mathcal{M}$, then $\mathcal{N}$ is a field.

(More generally, given an isomorphism $h : \mathcal{M} \cong \mathcal{N}$, an element $a \in M$ is invertible iff $h(a) \in N$ is, by Example 2.50 applied to h and h^{-1}.)

Exercise 2.58. An abelian group $\mathcal{M}$ is **torsion-free** if whenever $a \in M$ and $na := \overbrace{a + \cdots + a}^{n} = 0$ for some $n \geq 1$, then $a = 0$. For example, $\mathbb{Z}$ (with the usual $+$) is torsion-free, but $\mathbb{Z}/2\mathbb{Z}$ is not, since $1 + 1 = 0$ but $1 \neq 0$. Show that if two abelian groups are isomorphic, then one is torsion-free iff the other is.

Exercise 2.59. A **cycle** in a (simple undirected) graph $\mathcal{M} = (M, E^{\mathcal{M}})$ is a finite sequence of vertices $a_1, \ldots, a_n$ for some $n \geq 3$ such that $(a_1, a_2), (a_2, a_3), \ldots, (a_{n-1}, a_n), (a_n, a_1) \in E^{\mathcal{M}}$. A graph is a **forest** (also called **acyclic**) if it has no cycles. Show that if two graphs are isomorphic, then one is a forest iff the other is.

Exercise 2.60. An ordered field $\mathcal{M}$ is **Archimedean** if every element in it is $\leq \overbrace{1 + \cdots + 1}^{n}$ for some $n \in \mathbb{N}$. Show that if two ordered fields are isomorphic, then one is Archimedean iff the other is.

Exercise 2.61 (for those who know linear algebra)**.** Write down suitable first-order formulas to show that two isomorphic vector spaces have the same dimension.

As the last few examples suggest, the *finiteness* limitation on first-order formulas actually plays no role in the preceding preservation results (Propositions 2.49 and 2.55): roughly speaking, *any* expressible property of a structure, in however "higher-order" a logic one considers, will be preserved under isomorphisms. The following should give you a taste of such "higher-order" logics:

Exercise 2.62 (advanced)**.** An **infinitary signature** $\mathcal{A}$ is a set of function and relation symbols, where the arity of each symbol is an arbitrary set X;[4] as in first-order logic, we write $\mathcal{A}^X_{\text{fun}}, \mathcal{A}^X_{\text{rel}}$ to denote the X-ary function and relation symbols. The **infinitary $\mathcal{A}$-terms** are constructed inductively as follows:

- Every variable x is an $\mathcal{A}$-term.
- If $f \in \mathcal{A}^X_{\text{fun}}$ is an X-ary function symbol, and for each $x \in X$, we have an $\mathcal{A}$-term $\tau(x)$, then $f(\tau)$ is an $\mathcal{A}$-term.

The **infinitary $\mathcal{A}$-formulas** (also known as the **infinitary logic**[5] $\mathcal{L}_{\infty\infty}$) are constructed as follows:

- If $R \in \mathcal{A}^X_{\text{rel}}$ is an X-ary relation symbol, or the symbol $=$ when $X = \{1, 2\}$, and for each $x \in X$, we have an $\mathcal{A}$-term $\tau(x)$, then $R(\tau)$ is an **atomic $\mathcal{A}$-formula**.
- If X is an arbitrary set, and for each $x \in X$, we have an $\mathcal{A}$-formula ϕ_x, then $\bigwedge_{x\in X} \phi_x, \bigvee_{x\in X} \phi_x$ are $\mathcal{A}$-formulas.
- If ϕ is an $\mathcal{A}$-formula, then so is $\neg\phi$.
- If ϕ is an $\mathcal{A}$-formula, and X is an arbitrary set of variables, then $\exists X\, \phi$ is an $\mathcal{A}$-formula.

The **free variables** of an infinitary term/formula are defined the usual way; the only subtle case is

$$\text{FV}(\exists X\, \phi) := \text{FV}(\phi) \setminus X.$$

Let as usual $\mathcal{L}^X_{\text{term}}(\mathcal{A}), \mathcal{L}^X_{\text{form}}(\mathcal{A})$ denote the terms/formulas with free variables from X.

(a) Define the notion of $\mathcal{A}$-structure. [For instance, you should be able to turn $\mathbb{R}$ into an $\mathcal{A} = \{L\}$-structure, where L is an $\mathbb{N}$-ary relation symbol whose interpretation says that a given sequence $\alpha : \mathbb{N} \to \mathbb{R}$ has a limit.]

(b) Define the interpretation of $\mathcal{A}$-terms and $\mathcal{A}$-formulas in an $\mathcal{A}$-structure $\mathcal{M}$. [The $\exists$ case should be: for $\exists X\, \phi \in \mathcal{L}^Y_{\text{form}}(\mathcal{A})$,

$$\mathcal{M} \models_\alpha \exists X\, \phi \iff \exists \beta : X \to M \text{ s.t. } \mathcal{M} \models_{\alpha\langle\beta\rangle} \phi,$$

where $\alpha\langle\beta\rangle$ is what it looks like.]

(c) Prove Propositions 2.49 and 2.55 for infinitary formulas.

(d) Write down an infinitary formula ϕ with free variables $x, y_1, \ldots, y_n$, in the *finitary* signature $\mathcal{A}_{\mathsf{vec}}$, whose interpretation in a vector space says that x is a linear combination of $y_1, \ldots, y_n$.

(e) Write down an infinitary $\mathcal{A}_{\mathsf{vec}}$-formula with free variables $x_1, \ldots, x_n$ which says that $x_1, \ldots, x_n$ are linearly independent.

(f) Write down an infinitary $\mathcal{A}_{\mathsf{vec}}$-sentence which says that a vector space has dimension n.

(g) Write down an infinitary $\mathcal{A}_{\mathsf{poset}}$-sentence which says that every bounded sequence has a least upper bound (i.e., supremum).

(h) Write down a family of infinitary $\mathcal{A}_{\mathsf{poset}}$-sentences which say that every subset has a supremum. [Why does no single formula suffice?]

This also means that Corollary 2.56 gives a very weak criterion for axiomatizability: whereas we saw in propositional logic that axiomatizability was all about expressibility using *finite* formulas, the above tells us that isomorphism has nothing to do with the finite/infinite distinction. Indeed, the finiteness restriction on first-order logic means that certain homomorphisms which are *not* isomorphisms also preserve all formulas.

[4] If you prefer, you can assume that the arity is always an ordinal number α; only its cardinality matters.

[5] The first ∞ refers to the allowed arity of the $\bigwedge, \bigvee$'s; the second ∞ refers to the allowed arity of the $\exists$'s.

Definition 2.63. A homomorphism $h : \mathcal{M} \to \mathcal{N}$ is an **elementary embedding** if it preserves all first-order formulas.

Thus Proposition 2.55 says that all isomorphisms are elementary embeddings. Note that because an elementary embedding also has to preserve $\neg\phi$ for every ϕ, we can replace the $\Longrightarrow$ with $\Longleftrightarrow$ in the preservation condition:

$$\mathcal{M} \models_\alpha \phi \iff \mathcal{N} \models_{h\circ\alpha} \phi.$$

Of course, we already knew this for isomorphisms h, since h^{-1} is also an isomorphism; but general elementary embeddings need not be invertible. Intuitively, this is because the first-order formula ϕ, being finite, cannot really tell "how big" $\mathcal{M}, \mathcal{N}$ are, as long as they're "sufficiently infinite". While the full discussion of this idea will have to wait until we have the compactness theorem for first-order logic, you will prove the following simple case using only the tools we have right now:

Exercise 2.64 (HW7)**.** Let $h : M \to N$ be any function, regarded as an $\mathcal{A} = \varnothing$-homomorphism.

(a) Show that if h is an elementary embedding, then it must be injective.
(b) Show that if M is finite, then h is an elementary embedding iff it is bijective.
(c) Show that if M is infinite, then h is an elementary embedding iff it is injective.
(d) Conclude that no first-order $\varnothing$-theory can tell apart infinite sets of different cardinalities.

2.6. **Definability in structures.** We can also apply homomorphisms and isomorphisms to talk about a *different* way of connecting two first-order structures: instead of considering the same kind of structure on two different underlying sets, we can consider two different kinds of structure on the same underlying set.

Example 2.65. In Example 2.25, we defined an abelian group $\mathcal{M}$ to consist of operations $+, 0, -$. If you've taken a linear algebra course, you most likely defined a vector space there to consist only of the $+$ (as well as scalar multiplication) operations; the existence of 0 and additive inverses were instead enforced solely via axioms. These two notions are "equivalent", since $0, -$ are uniquely determined by $+$: the two structures $(\mathbb{Z}, +, 0, -)$ and $(\mathbb{Z}, +)$, say, are not *equal* (they aren't even structures over the same signature); but they still contain the same information in some sense.

Definition 2.66. Let $\mathcal{A}$ be a signature, $\mathcal{M}$ be an $\mathcal{A}$-structure. We say that an n-ary relation $R \subseteq M^n$ is **definable (from the $\mathcal{A}$-structure $\mathcal{M}$)**[6] if there is a formula $\phi \in \mathcal{L}_{\text{form}}^{\{x_1,\dots,x_n\}}(\mathcal{A})$ with n free variables such that

$$(a_1, \dots, a_n) \in R \iff \mathcal{M} \models_{x_1\mapsto a_1,\dots,x_n\mapsto a_n} \phi \quad \forall a_1, \dots, a_n \in M.$$

In other words, this says that $R \subseteq M^n$ agrees with $\phi^{\mathcal{M}}_{\{x_1,\dots,x_n\}} \subseteq M^{\{x_1,\dots,x_n\}}$, under the obvious bijection $M^n \cong M^{\{x_1,\dots,x_n\}}$.

We say that an n-ary function $f : M^n \to M$ is **definable** if the $(n+1)$-ary relation

$$\text{graph}(f) := \{(a_1, \dots, a_n, b) \in M^{n+1} \mid f(a_1, \dots, a_n) = b\}$$

is definable, i.e., there is a formula $\phi \in \mathcal{L}_{\text{form}}^{\{x_1,\dots,x_n,y\}}(\mathcal{A})$ such that

$$f(a_1, \dots, a_n) = b \iff \mathcal{M} \models_{x_1\mapsto a_1,\dots,x_n\mapsto a_n, y\mapsto b} \phi \quad \forall a_1, \dots, a_n, b \in M.$$

In particular, when $n \in 0$, this says that a constant $c \in M$ is definable iff there is $\phi \in \mathcal{L}_{\text{form}}^{\{y\}}(\mathcal{A})$ with a single free variable so that

$$c = b \iff \mathcal{M} \models_{y\mapsto b} \phi \quad \forall b \in M,$$

i.e., $\phi^{\mathcal{M}}_{\{y\}}$ is the single element c.

[6]The official term for this notion in model theory would be **definable from no parameters**.

Example 2.67. It would perhaps seem more natural to call a function $f : M^n \to M$ definable if it is defined by a *term* $t \in \mathcal{L}_{\text{term}}^{\{x_1,\dots,x_n\}}(\mathcal{A})$, i.e., if

$$f(a_1,\dots,a_n) = t^{\mathcal{M}}(x_1 \mapsto a_1,\ \dots,\ x_n \mapsto a_n) \quad \forall a_1,\dots,a_n \in M.$$

But this implies that (the graph of) f is defined, in the above sense, by the formula

$$\phi := (t = y) \in \mathcal{L}_{\text{form}}^{\{x_1,\dots,x_n,y\}}(\mathcal{A}).$$

Thus, definability of the graph of f is a more general notion than definability of f itself by a term. (The more restrictive notion of term-definability is useful for some purposes, but is *too* restrictive for many others, such as in the next few examples.)

Example 2.68. In the $\{+\}$-structure $\mathbb{N}$ (with the usual $+$), the binary relation $\leq$ is definable, via

$$\phi := \exists z\,(x + z = y) \in \mathcal{L}_{\text{form}}^{\{x,y\}}(\{+\}).$$

Indeed, for any $a, b \in \mathbb{N}$, we have

$$\begin{aligned} \mathbb{N} \models_{x\mapsto a, y\mapsto b} \exists z\,(x+z=y) &\iff \exists c \in \mathbb{N} \text{ s.t. } \mathbb{N} \models_{x\mapsto a, y\mapsto b, z\mapsto c} x + z = y \\ &\iff \exists c \in \mathbb{N} \text{ s.t. } a + c = b \\ &\iff a \leq b, \end{aligned}$$

since if $a + c = b$ then $b = a + c \geq a + 0 = a$, and conversely if $a \leq b$ then we can take $c := b - a$.

We can also define the constant $0 \in \mathbb{N}$, i.e., the singleton $\{0\} \subseteq \mathbb{N}$, via either

$$\phi := \forall y\,(y + x = y) \in \mathcal{L}_{\text{form}}^{\{x\}}(\{+\})$$

which says that x is the additive identity, or more simply via

$$\phi := (x + x = x)$$

(which has the benefit of being positive-existential; see Exercise 2.81 below).

Example 2.69. In the $\{+,\cdot\}$-structure $\mathbb{Z}$, $\leq$ is also definable. However, this is much less obvious than the above example, and requires use of the following fact from number theory as a black box:

Lagrange's four-square theorem. Any $n \in \mathbb{N}$ can be written as a sum of four perfect squares.

Given this, we may first define $\mathbb{N} \subseteq \mathbb{Z}$ via the formula

$$\phi := \exists a\,\exists b\,\exists c\,\exists d\,(z = a \cdot a + b \cdot b + c \cdot c + d \cdot d) \in \mathcal{L}_{\text{form}}^{\{z\}}(\{+,\cdot\});$$

indeed, if x is ≥ 0 then it can be written as a sum of four squares by Lagrange's four-square theorem, while conversely, clearly any sum of four squares is ≥ 0. Now to define $\leq$, we may use the formula ϕ from the previous example, but modified to ensure that the $\exists z$ only uses $z \in \mathbb{N}$:

$$\psi := \exists z\,(\text{“}z \in \mathbb{N}\text{”} \wedge (x + z = y)),$$

which, because $\mathbb{N}$ has already been defined by the formula ϕ, may be written as

$$:= \exists z\,(\phi \wedge (x + z = y)).$$

This example illustrates that in defining a relation or function, we may use other relations or functions that are already known to be definable. To state this more precisely, we have:

Proposition 2.70. Let $\mathcal{A} \subseteq \mathcal{B}$ be two signatures, and $\mathcal{M}$ be a $\mathcal{B}$-structure on an underlying set M. Suppose the interpretation of every symbol in $\mathcal{B}$ is definable from the $\mathcal{A}$-structure. Then every relation or function on M definable from the $\mathcal{B}$-structure is also definable from the $\mathcal{A}$-structure.

Proof. Since definability of a function on M amounts to definability of its graph, it suffices to consider definability of relations. Let $R \subseteq M^n$ be an n-ary relation definable from $\mathcal{B}$, say via the formula $\phi \in \mathcal{L}_{\text{form}}^{\{x_1,\dots,x_n\}}(\mathcal{B})$; we must get rid of the symbols in $\mathcal{B} \setminus \mathcal{A}$ from ϕ.

We first assume, for simplicity, that $\mathcal{A}_{\text{fun}} = \mathcal{B}_{\text{fun}}$, i.e., $\mathcal{B} \setminus \mathcal{A}$ consists only of relation symbols. In that case, the idea is to replace every relation symbol $S \in \mathcal{B} \setminus \mathcal{A}$ occurring in ϕ by the formula ψ defining S in $\mathcal{M}$ (this is analogous to the "formula substitutions" from HW2, only for first-order formulas). More precisely, we show the following by induction:

> For any $\mathcal{B}$-formula $\phi \in \mathcal{L}_{\text{form}}^X(\mathcal{B})$, there is an $\mathcal{A}$-formula $\phi' \in \mathcal{L}_{\text{form}}^X(\mathcal{A})$ with the same interpretation in $\mathcal{M}$.

In the base case where $\phi = S(t_1, \dots, t_n)$ is atomic, if $S \in \mathcal{A}$ then ϕ is already an $\mathcal{A}$-formula (since we are assuming $\mathcal{A}_{\text{fun}} = \mathcal{B}_{\text{fun}}$, so the terms $t_1, \dots, t_n$ only contain function symbols from $\mathcal{A}$), so we may take $\phi' := \phi$. Otherwise, $S \in \mathcal{B} \setminus \mathcal{A}$, so $S^{\mathcal{M}} \subseteq M^n$ is definable from the $\mathcal{A}$-structure, say by the $\mathcal{A}$-formula $\psi \in \mathcal{L}_{\text{form}}^{\{y_1,\dots,y_n\}}(\mathcal{A})$. Let

$$\phi' := \psi[y_1 \mapsto t_1, \dots, y_n \mapsto t_n] \in \mathcal{L}_{\text{form}}^X(\mathcal{A})$$

be the result of "substituting" the y_i's with the terms t_i's in ψ (this will be made precise in Section 3 below; see especially Corollary 3.26). This works, since for any $\alpha \in M^X$, we have

$$\begin{aligned}
\mathcal{M} \models_\alpha \phi &\iff (t_1^{\mathcal{M}}(\alpha), \dots, t_n^{\mathcal{M}}(\alpha)) \in S^{\mathcal{M}} && \text{since } \phi = S(t_1, \dots, t_n) \\
&\iff \mathcal{M} \models_{y_1 \mapsto t_1^{\mathcal{M}}(\alpha), \dots, y_n \mapsto t_n^{\mathcal{M}}(\alpha)} \psi && \text{since } \psi \text{ defines } S^{\mathcal{M}} \\
&\iff \mathcal{M} \models_\alpha \psi[y_1 \mapsto t_1, \dots, y_n \mapsto t_n] = \phi' && \text{by Lemma 4.41 below.}
\end{aligned}$$

This takes care of the base case; the inductive cases are all trivial, where we just apply whatever connective/quantifier to the formula(s) obtained from the IH.

Now suppose $\mathcal{B} \setminus \mathcal{A}$ (possibly) contains function symbols. Again, the only substantive case we need to consider is for atomic ϕ. First, consider ϕ of the form

$$(*) \qquad \phi = (t = y)$$

where $t \in \mathcal{L}_{\text{term}}^X(\mathcal{A})$ is a term; we show by induction on t that all such ϕ may be replaced by an $\mathcal{A}$-formula with the same interpretation in $\mathcal{M}$. If t is a variable, ϕ is already an $\mathcal{A}$-formula. Now suppose t begins with a function symbol $f \in \mathcal{B}_{\text{fun}}$:

$$\phi = (f(t_1, \dots, t_n) = y).$$

Note that such a formula is semantically equivalent to

$$(\dagger) \qquad \phi' := \exists y_1 \cdots \exists y_n \left((t_1 = y_1) \wedge \cdots \wedge (t_n = y_n) \wedge (f(y_1, \dots, y_n) = y) \right)$$

(where the $y_1, \dots, y_n$ are some new variables not appearing anywhere else). The last atomic subformula here, $f(y_1, \dots, y_n) = y$, defines the graph of f in $\mathcal{M}$, which we assumed to be definable from the $\mathcal{A}$-structure, say by an $\mathcal{A}$-formula $\psi \in \mathcal{L}_{\text{form}}^{\{y_1,\dots,y_n,z\}}(\mathcal{A})$; we may thus replace $f(y_1, \dots, y_n) = y$ with ψ in ϕ' without changing its interpretation in $\mathcal{M}$. Each of the other atomic subformulas $t_i = y_i$ in ϕ' may be replaced by an $\mathcal{A}$-formula with the same interpretation in $\mathcal{M}$ by the IH (since the $t_1, \dots, t_n$ are subterms of t). This completes the proof that atomic formulas of the form $(*)$ may be replaced. For a general atomic formula

$$\phi = S(t_1, \dots, t_n),$$

similarly to $(\dagger)$, this is semantically equivalent to

$$\phi' := \exists y_1 \cdots \exists y_n \left((t_1 = y_1) \wedge \cdots \wedge (t_n = y_n) \wedge S(y_1, \dots, y_n) \right);$$

the subformulas $t_i = y_i$ are of the form $(*)$, hence may be replaced, while the last subformula $S(y_1, \dots, y_n)$ may be replaced by a formula defining $S^{\mathcal{M}} \subseteq M^n$. □

Example 2.71. The following illustrates the procedure used in the second case (where $\mathcal{B} \setminus \mathcal{A}$ has function symbols) in the above proof. First, note that in the $\mathcal{A} = \{+\}$-structure $\mathbb{Z}$, the constant 0 is definable, by exactly the same formula as in Example 2.68:

$$\phi := (x + x = x).$$

Now in the $\mathcal{B} = \{+, 0\}$-structure $\mathbb{Z}$, negation $-$ may be defined via

$$\psi := (x + y = 0)$$

(i.e., "$x + y = 0 \iff -x = y$"). To show that $-$ is definable from only $+$, we need to eliminate the constant symbol 0 by "plugging in" the formula ϕ defining 0, yielding

$$\psi' := \exists z\, (\underbrace{(z + z = z)}_{\phi[x \mapsto z]} \wedge (x + y = z)).$$

An $\mathcal{A}_1$-structure $\mathcal{M}_1$ and an $\mathcal{A}_2$-structure $\mathcal{M}_2$ on the same underlying set M are called **interdefinable** if each function and relation in each structure is definable from the *other* structure. This concept formalizes the notion of "equivalence" from our motivating Example 2.65:

Example 2.72. The preceding example shows that $\mathbb{Z}$ equipped with the $\{+, 0, -\}$-structure is interdefinable with just the $\{+\}$-structure.

Similarly, $(\mathbb{Z}, +, \cdot)$ is interdefinable with the commutative ring $(\mathbb{Z}, +, 0, -, \cdot, 1)$ (and by Example 2.69, also with the ordered commutative ring $(\mathbb{Z}, +, 0, -, \cdot, 1, \leq)$).

Exercise 2.73. Show that $(\mathbb{R}, +, \cdot)$ is interdefinable with $\mathbb{R}$ equipped with:

(a) all of the ordered field structure;
(b) together with the cube root function;
(c) together with the absolute value function.

Show that $(\mathbb{R}, +, \cdot, \sin)$ is interdefinable with $\mathbb{R}$ equipped with all of the above, together with:

(d) the constant π;
(e) cos.

Exercise 2.74. Show that $\leq$ is definable in $(\mathbb{Q}, +, \cdot)$. [See lecture video.]

To show that something is *not* definable, we can apply the preservation results from the previous subsection: definable things have to be preserved under isomorphisms. More precisely:

Proposition 2.75. Let $h : \mathcal{M} \to \mathcal{N}$ be an $\mathcal{A}$-homomorphism between $\mathcal{A}$-structures.

(a) If a relation $R \subseteq M^n$ is **positive-existential definable**, i.e., defined by a positive-existential $\phi \in \mathcal{L}_{\text{form}}^{\{x_1,\ldots,x_n\}}(\mathcal{A})$, then h maps it into the relation $S \subseteq N^n$ defined by the *same* ϕ:
$$(a_1, \ldots, a_n) \in R \implies (h(a_1), \ldots, h(a_n)) \in S \;:\iff\; \mathcal{N} \models_{x_1 \mapsto h(a_1), \ldots, x_n \mapsto h(a_n)} \phi.$$

(b) If a function $f : M^n \to M$ is positive-existential definable by ϕ, *and* $\phi^{\mathcal{N}}$ *is also the graph of a function* $g : N^n \to N$, then h "maps f to g", i.e.,
$$h(f(a_1, \ldots, a_n)) = g(h(a_1), \ldots, h(a_n)).$$

If h is moreover an isomorphism, then these hold without the positive-existential restriction; and moreover, we may replace $\implies$ with $\iff$ in (a) (by considering h^{-1}):

$$(a_1, \ldots, a_n) \in R \iff (h(a_1), \ldots, h(a_n)) \in S.$$

Proof. (a) is immediate from Propositions 2.49 and 2.55; for (b), these same results yield

$$\begin{aligned} f(a_1, \ldots, a_n) = b &\iff \mathcal{M} \models_{x_1 \mapsto a_1, \ldots, x_n \mapsto a_n, y \mapsto b} \phi \\ &\implies \mathcal{N} \models_{x_1 \mapsto h(a_1), \ldots, x_n \mapsto h(a_n), y \mapsto h(b)} \phi \iff g(h(a_1), \ldots, h(a_n)) = h(b) \end{aligned}$$

which is clearly the same as the claimed equation (just take $b := f(a_1, \ldots, a_n)$). □

An isomorphism $h : \mathcal{M} \cong \mathcal{M}$ from a structure to itself is called an **automorphism**.

Corollary 2.76. If $R \subseteq M^n$ is definable, then it is preserved by every automorphism $h : \mathcal{M} \cong \mathcal{M}$:

$$(a_1, \ldots, a_n) \in R \iff (h(a_1), \ldots, h(a_n)) \in R.$$

Similarly, a definable function $f : M^n \to M$ is preserved by every automorphism.

Example 2.77. $\leq$ is *not* definable in $(\mathbb{Z}, +)$ (in contrast to Examples 2.68 and 2.69), since negation $- : \mathbb{Z} \to \mathbb{Z}$ is an abelian group automorphism ($-(a+b) = (-a) + (-b)$) but does not preserve $\leq$.

For the same reason, $1 \in \mathbb{Z}$ is *not* definable from $+$ (in contrast to Example 2.71).

It follows that $\cdot$ is not definable from $+$ either, or else 1, being definable from $\cdot$ (similarly to how 0 was defined from $+$ in Example 2.68), would be definable from $+$ as well (Proposition 2.70).

Exercise 2.78 (HW7)**.** The subset

$$\mathbb{Q}[\sqrt{2}] := \{p + q\sqrt{2} \mid p, q \in \mathbb{Q}\} \subseteq \mathbb{R}$$

is closed under $+, 0, -, \cdot, 1$ as well as reciprocals of nonzero elements, hence forms a subfield of $\mathbb{R}$, but has a non-identity automorphism

$$\begin{aligned} h : \mathbb{Q}[\sqrt{2}] &\longrightarrow \mathbb{Q}[\sqrt{2}] \\ p + q\sqrt{2} &\longmapsto p - q\sqrt{2} \end{aligned}$$

which does not preserve $\sqrt{2}$ or $\leq$, neither of which is therefore definable in $\mathbb{Q}[\sqrt{2}]$.

(Note the contrast with the two fields $\mathbb{Q} \subseteq \mathbb{Q}[\sqrt{2}] \subseteq \mathbb{R}$ sandwiching it, both of which have $\leq$ definable from the field structure by Exercises 2.73 and 2.74!)

Example 2.79. Similarly, the imaginary unit $i \in \mathbb{C}$ is not definable from the field structure, since complex conjugation $z \mapsto \overline{z}$ is a field automorphism that flips $\pm i$.

You may have heard before that $i = \sqrt{-1}$ is somehow "not uniquely defined", since there is "nothing distinguishing it" from the other square root $-i$ of -1: if we took all of math involving complex numbers, and replaced all the i's with $-i$'s, everything would still be valid. This "replacement" means applying the automorphism $z \mapsto \overline{z}$; and everything remains valid precisely because it is an automorphism, hence preserves everything we might want to say about $\mathbb{C}$.

(Of course, complex conjugation will only preserve things definable from structure that it preserves; if we add a constant symbol for i to the signature, then conjugation will definitely not preserve the formula $x = i$! The point is that the things we normally want to say about $\mathbb{C}$ only uses structure preserved by conjugation. Apart from the field structure, we might also care about the behavior of limits (i.e., the topology, such as when doing complex analysis) in $\mathbb{C}$; this can also be incorporated into the isomorphisms perspective, if we're willing to use infinitary logic (see Exercise 2.83 below).)

The "homomorphisms" part of Proposition 2.75 is also useful:

Example 2.80. A $\{+\}$-homomorphism $h : \mathcal{M} \to \mathcal{N}$ between abelian groups must also preserve $0, -$, i.e., be an abelian group homomorphism.

We proved this by hand in Proposition 2.36; but what the proof we gave there really showed was that $0, -$ are positive-existential definable from $+$ in any abelian group, by the exact same formulas as for $\mathbb{Z}$ in Example 2.71, hence are preserved by any $\{+\}$-homomorphism by Proposition 2.75(b).

Exercise 2.81. Recall from Exercise 2.37 that a $\{+\}$-homomorphism between abelian monoids need not preserve 0; you should be able to find such a counterexample whose domain is $\mathbb{N}$. How do you reconcile this with the fact that 0 is positive-existential definable from $+$ in $\mathbb{N}$ (Example 2.68)?

Exercise 2.82. Use definability to show that any *surjective* $\{+\}$-homomorphism between abelian monoids must preserve 0. [Recall HW6.]

We close this section by remarking that just as homomorphisms and isomorphisms are not inherently tied to *finitary* first-order logic (see discussion surrounding Exercise 2.62), definability can be naturally generalized to infinitary logic; moreover, the resulting generalized notion is what preservation under automorphisms *really* captures.

Exercise 2.83 (advanced)**.**

(a) Show that every rational in $\mathbb{R}$ (regarded as a constant) is definable from the field structure.
(b) Show that *every* element of $\mathbb{R}$ is definable, in *infinitary* logic (see Exercise 2.62), from the field structure. [Hint: recall that $\leq$ is definable from the field structure, by Exercise 2.73.]
(c) Conclude that the only field automorphism $h : \mathbb{R} \cong \mathbb{R}$ is the identity function.
(d) Show that *every* subset $A \subseteq \mathbb{R}$ is infinitary definable from the field structure. [Hint: use $\bigvee$.]
(e) Show that *every* n-ary relation $S \subseteq \mathbb{R}^n$ is infinitary definable from the field structure.

Exercise 2.84 (advanced)**.** Let $\mathcal{M}$ be an $\mathcal{A}$-structure.

(a) For any set X, give an infinitary $\varnothing$-formula $\phi \in \mathcal{L}^X_{\text{form}}(\varnothing)$ such that for any $\alpha : X \to M$,

$$\mathcal{M} \models_\alpha \phi \iff \alpha : X \to M \text{ is a bijection.}$$

[Hint: see HW7.]
(b) Give an infinitary $\mathcal{A}$-formula $\psi \in \mathcal{L}^M_{\text{form}}(\mathcal{A})$ such that for any $\alpha : M \to M$,

$$\mathcal{M} \models_\alpha \phi \iff \alpha : \mathcal{M} \to \mathcal{M} \text{ is an isomorphism.}$$

For an n-tuple $(a_1, \ldots, a_n) \in M^n$, its **orbit** is

$$[a_1, \ldots, a_n] := \{(h(a_1), \ldots, h(a_n)) \mid h : \mathcal{M} \cong \mathcal{M}\}.$$

(c) Show that every orbit is definable in infinitary logic.
(d) Show that an n-ary relation $R \subseteq M^n$ is preserved by every automorphism iff for every n-tuple $(a_1, \ldots, a_n) \in R$, the entire orbit $[a_1, \ldots, a_n] \subseteq R$.
(e) Conclude that $R \subseteq M^n$ is preserved by every automorphism iff it is definable in infinitary logic. [Hint: see parts (d) and (e) of the previous Exercise.]
(f) Extend this to X-ary relations $R \subseteq M^X$ (see Exercise 2.62), for arbitrary sets X.

3. Variable substitution

In order to perform nontrivial manipulations with the syntax of first-order logic, it is necessary to know how to substitute terms for free variables in formulas (as well as terms). In the proof of Proposition 2.70, we already saw an instance of this need. When we begin discussing first-order proofs in the following section, substitution will become absolutely essential; for instance, the rule for proving $\exists x\, \phi$ (see Definition 4.5) says that "it suffices to prove ϕ for some particular x", which formally means that we need to prove the result $\phi[x \mapsto t]$ of substituting some term t for x in ϕ.

Definition 3.1. Let $\mathcal{A}$ be a first-order signature. For two sets of variables X, Y, a **variable substitution from X to Y** is a function

$$\sigma : X \longrightarrow \mathcal{L}^Y_{\text{term}}(\mathcal{A}),$$

intuitively thought of as specifying, for each variable $x \in X$, a term $\sigma(x)$ with free variables from Y with which to replace x. Given such a σ, we define, for each term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, a term $t[\sigma] \in \mathcal{L}^Y_{\text{term}}(\mathcal{A})$, called the **substitution of σ into t**, inductively as follows:

$$\begin{aligned} x[\sigma] &:= \sigma(x) && \text{for } x \in X, \\ f(t_1, \ldots, t_n)[\sigma] &:= f(t_1[\sigma], \ldots, t_n[\sigma]) && \text{for } f \in \mathcal{A}^n_{\text{fun}} \text{ and } t_1, \ldots, t_n \in \mathcal{L}^X_{\text{term}}(\mathcal{A}). \end{aligned}$$

We then define, for each formula $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, a formula $\phi[\sigma] \in \mathcal{L}^Y_{\text{form}}(\mathcal{A})$, called the **substitution of σ into ϕ**, inductively as follows:

$$\begin{aligned}
R(t_1, \dots, t_n)[\sigma] &:= R(t_1[\sigma], \dots, t_n[\sigma]) \quad \text{for } R \in \mathcal{A}^n_{\text{rel}} \text{ (or } =), \ t_1, \dots, t_n \in \mathcal{L}^X_{\text{term}}(\mathcal{A}), \\
(\phi \wedge \psi)[\sigma] &:= \phi[\sigma] \wedge \psi[\sigma], \\
(\phi \vee \psi)[\sigma] &:= \phi[\sigma] \vee \psi[\sigma], \\
(\neg\phi)[\sigma] &:= \neg\phi[\sigma], \\
\top[\sigma] &:= \top, \\
\bot[\sigma] &:= \bot, \\
(\exists x\, \phi)[\sigma] &:= \exists x\, \phi[\sigma\langle x \mapsto x\rangle] \qquad \text{for } \phi \in \mathcal{L}^{X\cup\{x\}}_{\text{form}}(\mathcal{A}).
\end{aligned}$$

Here, as in Definition 2.11, $\sigma\langle x \mapsto x\rangle : X \cup \{x\} \to Y \cup \{x\}$ means σ extended with the assignment $x \mapsto x$, replacing any previous value of $\sigma(x)$.

It can be helpful, on an intuitive level, to think of variable substitution $\phi[\sigma]$ as analogous to semantic interpretation $\phi^{\mathcal{M}}(\alpha)$ as defined in Section 2.2, with σ playing the role of α; the difference is of course that we are "interpreting ϕ syntactically" instead of semantically. Analogously to Proposition 2.10, we have the following easy fact, whose proof is left as an

Exercise 3.2. $t[\sigma]$ and $\phi[\sigma]$ depend only on $\sigma(x)$ for those x occuring free in t, ϕ.

We have already met some other versions of substitution before: on HW2, you looked at substituting for atomic formulas (in propositional logic) instead of variables. The idea is mostly the same, except for the new complication caused by the presence of the variable-binding operation $\exists$:

Example 3.3. Consider the $\mathcal{A}_{\mathsf{ordfield}}$-formula

$$\phi := (x \le y) \wedge \exists y\, (x + y = z).$$

Informally speaking, the two occurrences of y don't refer to the same thing: the second occurrence is bound by the $\exists y$, hence is "inaccessible from the outside". This is reflected in the substitution

$$\begin{aligned}
\phi[x \mapsto 0,\, y \mapsto 1,\, z \mapsto z] &= (x \le y)[x \mapsto 0,\, y \mapsto 1,\, z \mapsto z] \wedge (\exists y\, (x + y = z))[x \mapsto 0,\, y \mapsto 1,\, z \mapsto z] \\
&= (x \le y)[x \mapsto 0,\, y \mapsto 1,\, z \mapsto z] \wedge \exists y\, (x + y = z)[x \mapsto 0,\, y \mapsto y,\, z \mapsto z] \\
&= (0 \le 1) \wedge \exists y\, (0 + y = z)
\end{aligned}$$

where in the middle step we used $(x \mapsto 0,\, y \mapsto 1,\, z \mapsto z)\langle y \mapsto y\rangle = (x \mapsto 0,\, y \mapsto y,\, z \mapsto z)$.

As in this example, it is common to want to substitute for only a few variables, while leaving all others unchanged. We therefore adopt the following convention: when we write $\phi[\sigma]$, we allow σ to be defined on only a subset of the free variables in ϕ, in which case we implicitly extend σ by the identity function on the remaining variables.

Example 3.4. With the same ϕ as in the previous example,

$$\begin{aligned}
\phi[x \mapsto y,\, y \mapsto z] &= (x \le y)[x \mapsto y,\, y \mapsto z] \wedge (\exists y\, (x + y = z))[x \mapsto y,\, y \mapsto z] \\
&= (x \le y)[x \mapsto y,\, y \mapsto z] \wedge \exists y\, (x + y = z)[x \mapsto y] \\
&= (y \le z) \wedge \exists y\, (y + y = z).
\end{aligned}$$

This example illustrates two important points. First, note that it is essential that we perform both substitutions $x \mapsto y$ and $y \mapsto z$ *simultaneously*. If we had first substituted $x \mapsto y$, then $y \mapsto z$, we would have ended up with $(y \le y)[y \mapsto z] = (z \le z)$ in the first clause, which is wrong.

Second, the second clause is *still* wrong: its meaning has been changed. For example, the original formula $\exists y\, (x + y = z)$, interpreted in $\mathbb{Z}$, should have been true for all values of x, z; but the new formula $\exists y\, (y + y = z)$ is no longer true when z is 1.

3.1. **Safe substitution.** In the preceding example, we say that the substitution of $\sigma := (x \mapsto y)$ into $\exists y\,(x + y = z)$ has **captured** the previously free occurrence of y in σ, turning it into the bound first occurrence of y in $\exists y\,(y + y = z)$. A little thought reveals the general cause of this:

Definition 3.5. We say that the substitution of $\sigma : X \to \mathcal{L}^Y_{\text{term}}(\mathcal{A})$ into the formula $\exists x\,\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ **captures** the variable x if $x \in \text{FV}(\sigma(y))$ for some $y \in \text{FV}(\exists x\,\phi) = \text{FV}(\phi) \setminus \{x\}$.

We say that the substitution of σ into a formula ϕ is **safe** if at no point during the substitution (including in any subformulas of ϕ) is any variable captured. Formally, this is defined by induction:

$$\begin{aligned}
R(t_1, \dots, t_n)[\sigma] &\text{ is always safe,} \\
(\phi \wedge \psi)[\sigma], (\phi \vee \psi)[\sigma] &\text{ are safe } :\iff \phi[\sigma], \psi[\sigma] \text{ are,} \\
(\neg\phi)[\sigma] &\text{ is safe } :\iff \phi[\sigma] \text{ is,} \\
\top[\sigma], \bot[\sigma] &\text{ are always safe,} \\
(\exists x\,\phi)[\sigma] &\text{ is safe } :\iff \text{it does not capture } x\text{, and } \phi[\sigma\langle x \mapsto x\rangle] \text{ is safe.}
\end{aligned}$$

(Substitution into a term is always considered safe, since terms do not bind variables.)

Example 3.6. The substitution

$$\begin{aligned}
&(\neg(x = 0) \to \exists y\,(x \cdot y = 1))[x \mapsto -y] \\
&= \neg(x = 0)[x \mapsto -y] \to (\exists y\,(x \cdot y = 1))[x \mapsto -y] \\
&= \neg(-y = 0) \to \exists y\,(x \cdot y = 1)[x \mapsto -y,\ y \mapsto y] \\
&= \neg(-y = 0) \to \exists y\,((-y) \cdot y = 1)
\end{aligned}$$

captures the variable y in the step $(\exists y\,(x \cdot y = 1))[x \mapsto -y]$, since the y in $\exists y$ occurs free in the assignment $x \mapsto -y$ to the variable $x \neq y$.

On the other hand, the substitution

$$\begin{aligned}
&(\forall x\,(\neg(x = 0) \to \exists y\,(x \cdot y = 1)))[x \mapsto -y] \\
&= \forall x\,(\neg(x = 0) \to \exists y\,(x \cdot y = 1))[x \mapsto x] \\
&= \cdots \\
&= \forall x\,(\neg(x = 0) \to \exists y\,(x \cdot y = 1))
\end{aligned}$$

is safe, since now y is no longer free in the assignment $x \mapsto x$.

We next state several technical lemmas expressing desirable properties of variable substitution. Note that in the formula cases, we need to assume the substitution is safe (see Exercise 3.10); this reflects the fact that safe substitutions are really the only meaningful ones.

Lemma 3.7 (HW8)**.** Let $\sigma : X \to \mathcal{L}^Y_{\text{term}}(\mathcal{A})$ be a variable substitution.

(a) For a term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we have $\text{FV}(t[\sigma]) = \bigcup_{x \in \text{FV}(t)} \text{FV}(\sigma(x))$.

(b) For a formula $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, we have $\text{FV}(\phi[\sigma]) = \bigcup_{x \in \text{FV}(\phi)} \text{FV}(\sigma(x))$, *assuming* $\phi[\sigma]$ *is safe.*

Lemma 3.8 (double substitution)**.** Let $\sigma : X \to \mathcal{L}^Y_{\text{term}}(\mathcal{A})$ and $\tau : Y \to \mathcal{L}^Z_{\text{term}}(\mathcal{A})$ be variable substitutions. We write $\sigma[\tau]$ for "τ composed with σ", i.e.,

$$\begin{aligned}
\sigma[\tau] : X &\longrightarrow \mathcal{L}^Z_{\text{term}}(\mathcal{A}) \\
x &\longmapsto \sigma(x)[\tau].
\end{aligned}$$

(a) For a term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we have $t[\sigma][\tau] = t[\sigma[\tau]]$.

(b) For a formula $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, we have $\phi[\sigma][\tau] = \phi[\sigma[\tau]]$, *assuming* $\phi[\sigma]$ *is safe.*

(c) In addition to the assumptions in (b), if $\phi[\sigma][\tau]$ is also safe, then so is $\phi[\sigma[\tau]]$.

Before we give the rather long and technical proof of this lemma, here is an

Example 3.9. According to part (b),

$$\begin{aligned}&(x+y\le z)[x\mapsto y,\,y\mapsto y,\,z\mapsto z][y\mapsto z]\\&=(x+y\le z)[x\mapsto y[y\mapsto z],y\mapsto y[y\mapsto z],z\mapsto z[y\mapsto z]]\\&=(x+y\le z)[x\mapsto z,y\mapsto z,z\mapsto z]\\&=(z+z\le z),\end{aligned}$$

which is indeed also what we get if we perform the two substitutions separately.

Warning: If, as per our convention (see after Example 3.3), we had abbreviated the first line as

$$(x+y\le z)[x\mapsto y][y\mapsto z],$$

we might have been tempted to erroneously apply (b) to rewrite this as

$$\begin{aligned}&=(x+y\le z)[x\mapsto z]\\&=(z+y\le z)\end{aligned}$$

which is wrong!

Proof of Lemma 3.8. (a) By induction on t.

- For $t=x\in X$, we have $t[\sigma][\tau]=\sigma(x)[\tau]=\sigma[\tau](x)=x[\sigma[\tau]]$.
- For $t=f(t_1,\dots,t_n)$, assuming the claim holds for $t_1,\dots,t_n$, we have

$$\begin{aligned}f(t_1,\dots,t_n)[\sigma][\tau]&=f(t_1[\sigma],\dots,t_n[\sigma])[\tau]\\&=f(t_1[\sigma][\tau],\dots,t_n[\sigma][\tau])\\&=f(t_1[\sigma[\tau]],\dots,t_n[\sigma[\tau]])\quad\text{by IH}\\&=f(t_1,\dots,t_n)[\sigma[\tau]].\end{aligned}$$

(b) By induction on ϕ. Most of the cases are similar to the inductive case in (a). The new case is

- For $\exists x\,\phi\in\mathcal{L}^X_{\text{form}}(\mathcal{A})$, assuming the claim holds for $\phi\in\mathcal{L}^{X\cup\{x\}}_{\text{form}}(\mathcal{A})$, we have

$$\begin{aligned}(\exists x\,\phi)[\sigma][\tau]&=(\exists x\,\phi[\sigma\langle x\mapsto x\rangle])[\tau]\\&=\exists x\,\phi[\sigma\langle x\mapsto x\rangle][\tau\langle x\mapsto x\rangle]\\(*)\qquad&=\exists x\,\phi[\sigma\langle x\mapsto x\rangle[\tau\langle x\mapsto x\rangle]];\end{aligned}$$

we want to show that this is

$$(\dagger)\qquad=\exists x\,\phi[\sigma[\tau]\langle x\mapsto x\rangle]=(\exists x\,\phi)[\sigma[\tau]].$$

We have

$$\begin{aligned}\sigma[\tau]\langle x\mapsto x\rangle:X\cup\{x\}&\longrightarrow\mathcal{L}^{Z\cup\{x\}}_{\text{form}}(\mathcal{A})\\x&\longmapsto x,\\X\setminus\{x\}\ni y&\longmapsto\sigma[\tau](y)=\sigma(y)[\tau],\end{aligned}$$

while

$$\begin{aligned}\sigma\langle x\mapsto x\rangle[\tau\langle x\mapsto x\rangle]:X\cup\{x\}&\longrightarrow\mathcal{L}^{Z\cup\{x\}}_{\text{form}}(\mathcal{A})\\x&\longmapsto\sigma\langle x\mapsto x\rangle(x)[\tau\langle x\mapsto x\rangle]=x[\tau\langle x\mapsto x\rangle]=x,\\X\setminus\{x\}\ni y&\longmapsto\sigma\langle x\mapsto x\rangle(y)[\tau\langle x\mapsto x\rangle]=\sigma(y)[\tau\langle x\mapsto x\rangle];\end{aligned}$$

for all $y\in\mathrm{FV}(\phi)$, this last term $\sigma(y)[\tau\langle x\mapsto x\rangle]$ is the same as $\sigma(y)[\tau]$, since $x\notin\mathrm{FV}(\sigma(y))$, since $y\neq x$ and the substitution $(\exists x\,\phi)[\sigma]$ is assumed to be safe. Thus $(*)=(\dagger)$, since the two substitutions agree on all those variables which actually occur free in ϕ.

(c) By induction on ϕ.

- For ϕ atomic, or $\phi = \top, \bot$, $\phi[\sigma[\tau]]$ is always safe.
- If claim holds for ϕ, ψ, and $(\phi \wedge \psi)[\sigma][\tau] = (\phi[\sigma] \wedge \psi[\sigma])[\tau]$ is safe, that by definition means $\phi[\sigma][\tau], \psi[\sigma][\tau]$ are safe, whence by the IH, so are $\phi[\sigma[\tau]], \psi[\sigma[\tau]]$, hence so is $(\phi \wedge \psi)[\sigma[\tau]]$. Similarly for $\vee, \neg$.
- Finally, suppose the claim holds for $\phi \in \mathcal{L}^{X\cup\{x\}}_{\text{form}}(\mathcal{A})$, and $(\exists x\, \phi)[\sigma][\tau] = (\exists x\, \phi[\sigma\langle x \mapsto x\rangle])[\tau]$ is safe. That by definition means $\phi[\sigma\langle x \mapsto x\rangle][\tau\langle x \mapsto x\rangle]$ is safe, which by the IH implies that $\phi[\sigma[\tau]\langle x \mapsto x\rangle]$ is safe by the computation in (b), *and also* that $(\exists x\, \phi[\sigma\langle x \mapsto x\rangle])[\tau]$ does not capture x, which means that

$$(*) \qquad x \notin \mathrm{FV}(\tau(y)) \quad \text{for any } y \in \mathrm{FV}(\phi[\sigma\langle x \mapsto x\rangle]) \setminus \{x\}.$$

We must show that this implies that $(\exists x\, \phi)[\sigma[\tau]]$ does not capture x, i.e.,

$$(\dagger) \qquad x \notin \mathrm{FV}(\sigma[\tau](z)) = \mathrm{FV}(\sigma(z)[\tau]) \quad \text{for any } z \in \mathrm{FV}(\phi) \setminus \{x\}.$$

To see this: by Lemma 3.7,

$$\mathrm{FV}(\sigma(z)[\tau]) = \textstyle\bigcup_{y\in\mathrm{FV}(\sigma(z))} \mathrm{FV}(\tau(y));$$

thus we must show that $x \notin \mathrm{FV}(\tau(y))$ for all $y \in \mathrm{FV}(\sigma(z))$. By $(*)$, it suffices to show

$$y \in \mathrm{FV}(\sigma(z)) \implies y \in \mathrm{FV}(\phi[\sigma\langle x \mapsto x\rangle]) \setminus \{x\}.$$

Indeed, for $y \in \mathrm{FV}(\sigma(z))$, we have $y \neq x$, or else $(\exists x\, \phi)[\sigma]$ would capture x since $z \in \mathrm{FV}(\phi) \setminus \{x\}$ (by the assumption in $(\dagger)$), contradicting the safety assumption in (b). And since $y \in \mathrm{FV}(\sigma(z)) = \mathrm{FV}(\sigma\langle x \mapsto x\rangle(z))$ (since $z \neq x$) and $z \in \mathrm{FV}(\phi)$, we get $y \in \mathrm{FV}(\phi[\sigma\langle x \mapsto x\rangle]) = \bigcup_{z\in\mathrm{FV}(\phi)} \mathrm{FV}(\sigma\langle x \mapsto x\rangle(z))$ again by Lemma 3.7. □

Exercise 3.10 (HW8)**.** Find counterexamples to the relevant parts of the above two lemmas when the safety assumptions are dropped.

Corollary 3.11. Let $\sigma : X \cong Y$ be a *bijection* between two sets of variables.

(a) For $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we have $t[\sigma][\sigma^{-1}] = t$, thus yielding a bijection

$$\begin{aligned} \mathcal{L}^X_{\text{term}}(\mathcal{A}) &\cong \mathcal{L}^Y_{\text{term}}(\mathcal{A}) \\ t &\mapsto t[\sigma] \\ s[\sigma^{-1}] &\mapsfrom s. \end{aligned}$$

(b) For $\phi \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, if $\phi[\sigma]$ is safe, then so is $\phi[\sigma][\sigma^{-1}] = \phi$. (Thus, we get a bijection between the subsets of $\mathcal{L}^X_{\text{form}}(\mathcal{A}), \mathcal{L}^Y_{\text{form}}(\mathcal{A})$ of those ϕ for which $\phi[\sigma], \phi[\sigma^{-1}]$ respectively are safe.)

Proof. The equations $t[\sigma][\sigma^{-1}] = t[\sigma[\sigma^{-1}]] = t[\sigma^{-1} \circ \sigma] = t$, and similarly $\phi[\sigma][\sigma^{-1}] = \phi$, follow immediately from the double substitution Lemma 3.8. It remains only to verify in (b) that if $\phi[\sigma]$ is safe, then so is $\phi[\sigma][\sigma^{-1}]$. This is again by induction on ϕ, where the only nontrivial case is for $\exists x\, \phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$. To say that

$$(\exists x\, \phi)[\sigma] = \exists x\, \phi[\sigma\langle x \mapsto x\rangle]$$

is safe means that $\phi[\sigma\langle x \mapsto x\rangle]$ is safe, and also $x \neq \sigma(y)$ for any $y \in \mathrm{FV}(\phi) \setminus \{x\}$. It follows that $\sigma\langle x \mapsto x\rangle|\mathrm{FV}(\phi)$ is still an injection, hence a bijection with its image, which by Lemma 3.7 is

$$\sigma\langle x \mapsto x\rangle(\mathrm{FV}(\phi)) = \mathrm{FV}(\phi[\sigma\langle x \mapsto x\rangle]).$$

Thus for $y \in \mathrm{FV}(\phi[\sigma\langle x \mapsto x\rangle]) \setminus \{x\} = \sigma(\mathrm{FV}(\phi) \setminus \{x\})$, we have $\sigma^{-1}(y) \in \mathrm{FV}(\phi) \setminus \{x\}$, whence

$$(\exists x\, \phi)[\sigma][\sigma^{-1}] = (\exists x\, \phi[\sigma\langle x \mapsto x\rangle])[\sigma^{-1}] = \exists x\, \phi[\sigma\langle x \mapsto x\rangle][\sigma^{-1}\langle x \mapsto x\rangle]$$

does not capture x; and the inner substitution is also safe by the IH, since $\sigma\langle x \mapsto x\rangle$ and $\sigma^{-1}\langle x \mapsto x\rangle$ become inverses of each other when restricted to $\mathrm{FV}(\phi)$ and $\mathrm{FV}(\phi[\sigma\langle x \mapsto x\rangle])$. □

3.2. **α-equivalence.** The solution to variable capture as in Example 3.4 is familiar from informal mathematical practice: in order to substitute $x \mapsto y$ into $\exists y\,(x + y = z)$ without breaking its meaning, we must first replace the bound variable y with a different variable, say $\exists w\,(x + w = z)$. The resulting formula is, as usual, not syntactically *equal* to the original formula; but it is equivalent in a very fine syntactic sense (much finer than provable or semantic equivalence, say). In order to formalize this notion of equivalence, we need several steps:

Definition 3.12. We say that an existential formula $\exists x\,\phi$ is **immediately α-equivalent**,[7] denoted $\sim_\alpha$, to the result of changing x to a new variable y *in both the* $\exists x$ *and in* ϕ via a safe substitution:

$$\exists x\,\phi \sim_\alpha \exists y\,\phi[x \mapsto y] \quad \text{where } y \notin \mathrm{FV}(\phi) \cup \{x\} \text{ and } \phi[x \mapsto y] \text{ is safe.}$$

We say that two formulas ϕ, ψ are **one-step α-equivalent**, denoted

$$\phi \approx_\alpha \psi,$$

if they are immediately α-equivalent at a single subformula position; more precisely, $\approx_\alpha$ is the binary relation on formulas defined inductively as follows:

$$\begin{aligned}
\psi \sim_\alpha \phi &\implies \phi \approx_\alpha \psi,\\
\phi \approx_\alpha \psi &\implies \phi \wedge \theta \approx_\alpha \psi \wedge \theta,\\
\phi \approx_\alpha \psi &\implies \theta \wedge \phi \approx_\alpha \theta \wedge \psi,\\
\phi \approx_\alpha \psi &\implies \phi \vee \theta \approx_\alpha \psi \vee \theta,\\
\phi \approx_\alpha \psi &\implies \theta \vee \phi \approx_\alpha \theta \vee \psi,\\
\phi \approx_\alpha \psi &\implies \neg\phi \approx_\alpha \neg\psi,\\
\phi \approx_\alpha \psi &\implies \exists x\,\phi \approx_\alpha \exists x\,\psi.
\end{aligned}$$

Finally, we say that two formulas are **α-equivalent** if they are linked by a finite sequence of one-step α-equivalences, denoted

$$\phi \equiv_\alpha \psi \;:\Longleftrightarrow\; \exists \phi_0, \phi_1, \ldots, \phi_n \text{ s.t. } \phi = \phi_0 \approx_\alpha \phi_1 \approx_\alpha \cdots \approx_\alpha \phi_n = \psi.$$

(Thus, when $n = 1$, this just means $\phi \approx_\alpha \psi$; when $n = 0$, it means $\phi = \psi$.)

Example 3.13. Recalling the formula from Example 3.4, we have

$$\exists y\,(x + y = z) \sim_\alpha \exists w\,(x + y = z)[y \mapsto w] = \exists w\,(x + w = z),$$

since the substitution $(x + y = z)[y \mapsto w]$ is clearly safe. Thus

$$(x \le y) \wedge \exists y\,(x + y = z) \approx_\alpha (x \le y) \wedge \exists w\,(x + w = z)$$

(hence also $\equiv_\alpha$). However,

$$\exists y\,(x + y = z) \not\sim_\alpha \exists x\,(x + y = z)[y \mapsto x] = \exists x\,(x + x = z)$$

since $x \in \mathrm{FV}(x + y = z)$.

Example 3.14. We have

$$\forall x\,\exists y\,(x + y = 0) \equiv_\alpha \forall y\,\exists x\,(y + x = 0).$$

Since there are two quantifiers whose variables changed, these formulas cannot be $\approx_\alpha$; we need at least two steps. But two steps are not enough: we cannot immediately change the y in $\exists y\,(x + y = 0)$

[7] The "α" here is part of the conventional terminology, and does not denote a variable of any kind; in particular, it is not a variable assignment. Note also that $\exists y\,\phi[x \mapsto y]$ is itself *not* the result of any substitution into $\exists x\,\phi$.

to x, since $x \in \mathrm{FV}(x + y = 0)$. And we cannot immediately change the x in $\forall x$ to y either, since the substitution $(\exists y\,(x + y = 0))[x \mapsto y]$ is not safe. Instead, we need to go through a third variable:

$$\exists y\,(x + y = 0) \sim_\alpha \exists z\,(x + z = 0),$$

whence

$$\text{(a)} \qquad \forall x\,\exists y\,(x + y = 0) \approx_\alpha \forall x\,\exists z\,(x + z = 0);$$

and now the substitution $(\exists z\,(x + z = 0))[x \mapsto y]$ is safe, so

$$\text{(b)} \qquad \forall x\,\exists z\,(x + z = 0) \sim_\alpha \forall y\,\exists z\,(y + z = 0)$$

(hence also $\approx_\alpha$); finally, for similar reasons as in (a),

$$\text{(c)} \qquad \forall y\,\exists z\,(y + z = 0) \approx_\alpha \forall y\,\exists x\,(y + x = 0),$$

whence by chaining together (a), (b), and (c) we get the desired $\equiv_\alpha$.

Exercise 3.15. Show that the following formulas are α-equivalent, in as few steps as possible:

$$\forall x\,((\exists x\,(x \le y)) \to (\exists y\,(x \le y))),$$
$$\forall y\,((\exists y\,(y \le x)) \to (\exists x\,(y \le x))).$$

Explain why fewer steps are not possible.

Proposition 3.16. $\equiv_\alpha$ is an equivalence relation on formulas.

Proof. The only nontrivial part is to show that $\sim_\alpha$ is symmetric. Suppose $\exists x\,\phi \sim_\alpha \exists y\,\phi[x \mapsto y]$, where $y \notin \mathrm{FV}(\phi) \cup \{x\}$ and $\phi[x \mapsto y]$ is safe. Let $\sigma : \mathrm{FV}(\phi) \cup \{x\} \to (\mathrm{FV}(\phi) \setminus \{x\}) \cup \{y\}$ be the bijection mapping $x \mapsto y$ and every other variable to itself, so that $\phi[x \mapsto y] = \phi[\sigma]$. Then $\phi[x \mapsto y][y \mapsto x] = \phi[\sigma][\sigma^{-1}] = \phi$ is a safe substitution by Corollary 3.11, and $x \notin \mathrm{FV}(\phi[\sigma])$ since $\mathrm{FV}(\phi[\sigma]) \subseteq (\mathrm{FV}(\phi) \setminus \{x\}) \cup \{y\}$, which shows $\exists y\,\phi[x \mapsto y] \sim_\alpha \exists x\,\phi[x \mapsto y][y \mapsto x] = \exists x\,\phi$.

It now follows by a trivial induction that $\approx_\alpha$ is also symmetric. Thus, $\equiv_\alpha$ is symmetric since we may reverse the finite string of $\approx_\alpha$'s, transitive since we may join together two such finite strings, and reflexive since we allowed strings of length $n = 0$. □

We now show that "everything important you can do with formulas respects $\equiv_\alpha$". In other words, every major operation or property of formulas we have discussed thus far descends to the quotient set $\mathcal{L}_{\mathrm{form}}(\mathcal{A})/\equiv_\alpha$, and so can be thought of as operating on α-equivalence classes of formulas.[8]

Proposition 3.17. Logical connectives and quantifiers preserve $\equiv_\alpha$: if $\phi \equiv_\alpha \psi$, then

$$\phi \wedge \theta \equiv_\alpha \psi \wedge \theta,$$
$$\theta \wedge \phi \equiv_\alpha \theta \wedge \psi,$$
$$\phi \vee \theta \equiv_\alpha \psi \vee \theta,$$
$$\theta \vee \phi \equiv_\alpha \theta \vee \psi,$$
$$\neg\phi \equiv_\alpha \neg\psi,$$
$$\exists x\,\phi \equiv_\alpha \exists x\,\psi.$$

(Thus, for instance, if $\phi \equiv_\alpha \phi'$ and $\psi \equiv_\alpha \psi'$, then $\phi \wedge \psi \equiv_\alpha \phi' \wedge \psi \equiv_\alpha \phi' \wedge \psi'$.)

Proof. By Definition 3.12, logical connectives and quantifiers preserve $\approx_\alpha$; now apply this to each of the $\approx_\alpha$'s in the finite string $\phi = \phi_0 \approx_\alpha \phi_1 \approx_\alpha \cdots \approx_\alpha \phi_n = \psi$ witnessing $\equiv_\alpha$. For instance, to show $\phi \wedge \theta \equiv_\alpha \psi \wedge \theta$: we have

$$\phi \wedge \theta = \phi_0 \wedge \theta \approx_\alpha \phi_1 \wedge \theta \approx_\alpha \cdots \approx_\alpha \phi_n \wedge \theta = \psi \wedge \theta.$$ □

[8] In more advanced logic textbooks, it is common to simply sweep α-equivalence under the rug and identify formulas up to α-equivalence to begin with.

Proposition 3.18. If $\phi \equiv_\alpha \psi$, then $\mathrm{FV}(\phi) = \mathrm{FV}(\psi)$.

Proof. First, suppose $\exists x\, \phi \sim_\alpha \exists y\, \phi[x \mapsto y]$, where $y \notin \mathrm{FV}(\phi) \cup \{x\}$ and $\phi[x \mapsto y]$ is safe. Then

$$\begin{aligned}\mathrm{FV}(\exists y\, \phi[x \mapsto y]) &= \mathrm{FV}(\phi[x \mapsto y]) \setminus \{y\} \\ &= \mathrm{FV}(\phi) \setminus \{x\} \quad \text{by Lemma 3.7} \\ &= \mathrm{FV}(\exists x\, \phi).\end{aligned}$$

Now by a trivial induction, if $\phi \approx_\alpha \psi$, then $\mathrm{FV}(\phi) = \mathrm{FV}(\psi)$; the claim for $\equiv_\alpha$ follows. □

We defer the proof of the following until Lemma 4.42, since it involves semantics:

Proposition 3.19. α-equivalent formulas are semantically equivalent.

The last important operation on formulas we have discussed is substitution itself:

Proposition 3.20 (safe substitution preserves α-equivalence)**.** If $\phi \equiv_\alpha \psi \in \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$, and $\sigma : X \to \mathcal{L}^Y_{\mathrm{term}}(\mathcal{A})$ is a substitution such that both $\phi[\sigma], \psi[\sigma]$ are safe, then $\phi[\sigma] \equiv_\alpha \psi[\sigma]$.

Exercise 3.21. Try to prove this directly, first for $\sim_\alpha$, then $\approx_\alpha$, then $\equiv_\alpha$, to see what goes wrong.

The correct proof of Proposition 3.20 is rather long, and deferred to the end of this subsection (see around Proposition 3.28). Before giving it, we revisit our original motivation for introducing α-equivalence: we prove that indeed, we may change any formula to an α-equivalent copy that avoids all variable clashes. We may phrase this via the following intuitively obvious notion, which we have avoided defining rigorously until now (since free variables are much more important):

Exercise 3.22.

(a) Inductively define the set of **bound variables** $\mathrm{BV}(\phi)$ of a formula ϕ.
[For the answer, see the proof of Proposition 3.24 below.]

(b) From Definition 3.5, it is clear that the only variables that might be captured during a substitution $\phi[\sigma]$ are those which appear bound in ϕ but free in $\sigma(y)$ for some $y \in \mathrm{FV}(\phi)$. Prove this rigorously by induction on ϕ.

(c) Prove that for any variable substitution σ (even if unsafe), $\mathrm{BV}(\phi[\sigma]) = \mathrm{BV}(\phi)$.

Convention 3.23. Up to now, we have not said much about the alphabet $\mathcal{V}$ from which variables are drawn, aside from assuming it as given in Definition 1.8.

We henceforth assume that $\mathcal{V}$ has infinitely many variables outside of any given set X of variables under consideration. For instance, for any formula ϕ, or set of formulas (i.e., a theory), there are infinitely many variables not appearing either free or bound in those formula(s). This assumption is justified, because if we ever find that we have exhausted the set of all possible variables, we can always go back and pretend we had started with a bigger $\mathcal{V}' \supsetneq \mathcal{V}$ to begin with.[9]

Proposition 3.24. For any formula ϕ, and any infinite set of variables $X \subseteq \mathcal{V}$, there is a $\psi \equiv_\alpha \phi$ such that $\mathrm{BV}(\psi) \subseteq X$.

Intuitively, this says that any formula may be rewritten to only use bound variables from a "safe" set X. Before giving the proof, we first deduce its most important consequence:

Corollary 3.25. For any $\phi \in \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$ and $\sigma : X \to \mathcal{L}^Y_{\mathrm{term}}(\mathcal{A})$, there is $\psi \equiv_\alpha \phi$ making $\psi[\sigma]$ safe.

Proof. By Convention 3.23, there are infinitely many variables not appearing in $\sigma(y)$ for any $y \in \mathrm{FV}(\phi)$. By Proposition 3.24, there is $\psi \equiv_\alpha \phi$ whose bound variables are among these. By Exercise 3.22(b), this ensures that $\psi[\sigma]$ is safe. □

[9] The cleanest set-theoretic solution is to simply forget about the fixed set $\mathcal{V}$, and allow any mathematical object whatsoever to be a "variable"; this convention then reduces to the fact that there is no set of all mathematical objects.

Corollary 3.26. *Safe* substitution is a well-defined operation on α-equivalence classes of formulas.

Proof. It is always defined by Corollary 3.25, and the result is unique by Proposition 3.20. □

Proof of Proposition 3.24. By induction on ϕ.

- For ϕ atomic or $\top, \bot$, $\mathrm{BV}(\phi) = \varnothing$ (by your definition from Exercise 3.22(a)), so put $\psi := \phi$.
- If the claim holds for ϕ, ψ, then to prove it for $\phi \wedge \psi$, let $\phi' \equiv_\alpha \phi$ and $\psi' \equiv_\alpha \psi$ with $\mathrm{BV}(\phi'), \mathrm{BV}(\psi') \subseteq X$; then (by your definition from Exercise 3.22(a))
$$\mathrm{BV}(\phi' \wedge \psi') = \mathrm{BV}(\phi') \cup \mathrm{BV}(\psi') \subseteq X,$$
and by Proposition 3.17, $\phi' \wedge \psi' \equiv_\alpha \phi \wedge \psi$. The other connectives are similar.
- Finally, suppose the claim holds for ϕ; we prove it for $\exists x\, \phi$. Pick some $y \in X \setminus (\mathrm{FV}(\phi) \cup \{x\})$, using that X is infinite while $\mathrm{FV}(\phi)$ is finite. By the IH, there is $\phi' \equiv_\alpha \phi$ with
$$(*) \qquad \mathrm{BV}(\phi') \subseteq X \setminus \{y\}.$$
Then by Exercise 3.22(b), $\phi'[x \mapsto y]$ is safe, so since $y \in X \setminus (\mathrm{FV}(\phi) \cup \{x\})$,
$$\exists y\, \phi'[x \mapsto y] \sim_\alpha \exists x\, \phi' \equiv_\alpha \exists x\, \phi \quad \text{by Proposition 3.17.}$$
Finally, by your definition from Exercise 3.22(a),
$$\begin{aligned} \mathrm{BV}(\exists y\, \phi'[x \mapsto y]) &= \mathrm{BV}(\phi'[x \mapsto y]) \cup \{y\} \\ &= \mathrm{BV}(\phi') \cup \{y\} && \text{by Exercise 3.22(c)} \\ &\subseteq X && \text{by } (*). \end{aligned}$$
□

Exercise 3.27. A formula ϕ satisfies the **Barendregt variable convention** if the variables bound by different quantifiers in it are all distinct from each other and from all free variables.

(a) Define what this means precisely.
(b) Prove that any formula ϕ is α-equivalent to one satisfying the Barendregt variable convention.

We conclude this section with the proof of Proposition 3.20. Intuitively speaking, the reason the proof is tricky (which you might have seen in Exercise 3.21) is that the definition of $\equiv_\alpha$ allows us to change variables in a highly disorganized fashion, jumping between different subformulas (see Example 3.14). The following "converse" to Proposition 3.17 tells us that we may always rearrange the sequence of $\approx_\alpha$'s so as to reflect the inductive structure of the formulas themselves:

Proposition 3.28 (structural characterization of $\equiv_\alpha$)**.** Let $\phi \equiv_\alpha \psi$.

(a) If ϕ is atomic, $\top$, or $\bot$, then $\phi = \psi$.
(b) If ϕ is a conjunction, then so is ψ, and we have
$$\phi = \phi' \wedge \phi'' \equiv_\alpha \psi' \wedge \psi'' = \psi$$
for some $\phi' \equiv_\alpha \psi'$ and $\phi'' \equiv_\alpha \psi''$.
(c) If ϕ is a disjunction, then so is ψ, and we have
$$\phi = \phi' \vee \phi'' \equiv_\alpha \psi' \vee \psi'' = \psi$$
for some $\phi' \equiv_\alpha \psi'$ and $\phi'' \equiv_\alpha \psi''$.
(d) If ϕ is a negation, then so is ψ, and we have
$$\phi = \neg\phi' \equiv_\alpha \neg\psi' = \psi$$
for some $\phi' \equiv_\alpha \psi'$.
(e) If ϕ is an existential, then so is ψ, and we have
$$\phi = \exists x\, \phi' \sim_\alpha \exists z\, \phi'[x \mapsto z] \equiv_\alpha \exists z\, \psi'[y \mapsto z] \sim_\alpha \exists y\, \psi' = \psi$$
for some ϕ', ψ', such that $\phi'[x \mapsto z] \equiv_\alpha \psi'[y \mapsto z]$ for *any* variable z witnessing both of the outer $\sim_\alpha$'s (i.e., $z \notin \mathrm{FV}(\phi') \cup \mathrm{FV}(\psi') \cup \{x, y\}$ and $\phi'[x \mapsto z], \psi'[y \mapsto z]$ are safe).

We will give the proofs of Propositions 3.20 and 3.28 simultaneously. (It would have been possible to phrase the proof in terms of Proposition 3.20 only, but the proof would've implicitly proved Proposition 3.28 along the way, so it seems clearer to explicitly state the latter result, which can be useful in its own right.)

The proof will be by induction on formulas; however, in the inductive case for $\exists x\,\phi$, we will need to refer to the IH not just for ϕ but for a substituted copy of it. Thus, this is not really an induction on formulas, since we are reducing not to a subformula of the original formula, but rather to a slightly different formula with the same "size". Here "size" can be any of several numerical measures you've seen before (e.g., on HW1), such as

Exercise 3.29.

(a) Inductively define the **height** $\mathrm{HT}(\phi)$ of a first-order formula. [See Quiz 1. For the following proof, what's important is that a subformula has smaller height than a formula containing it; if you want, you can let all atomic formulas have height 0, ignoring the sizes of terms.]

(b) Prove that for any variable substitution $\sigma : X \to Y$ (you could also allow substituting terms, if you ignored the sizes of terms in (a)), we have $\mathrm{HT}(\phi[\sigma]) = \mathrm{HT}(\phi)$ (even if unsafe).

(c) Prove that if $\phi \equiv_\alpha \psi$, then $\mathrm{HT}(\phi) = \mathrm{HT}(\psi)$.

Proof of Propositions 3.20 and 3.28. Since $\phi \equiv_\alpha \psi$, let

$$\phi = \phi_0 \approx_\alpha \phi_1 \approx_\alpha \cdots \approx_\alpha \phi_n = \psi.$$

In Proposition 3.28(e), we will first prove the weaker statement where z may be any variable outside of *some* finite set (while (e) says it is enough to take z not appearing either free or bound in ϕ', ψ', so that $\phi'[x \mapsto z], \psi'[y \mapsto z]$ are safe by Exercise 3.22(b)). We proceed by induction on $\mathrm{HT}(\phi)$.

- If ϕ is atomic, $\top$, or $\bot$, there is no clause in the definition of $\approx_\alpha$ which yields $\phi = \phi_0 \approx_\alpha \phi_1$; thus the above sequence must have length $n = 0$, i.e., $\phi = \psi$, whence clearly $\phi[\sigma] = \psi[\sigma]$.
- If $\phi = \phi' \wedge \phi''$, then by considering the possibilities for $\phi = \phi_0 \approx_\alpha \phi_1$, we must have $\phi_1 = \phi_1' \wedge \phi_1''$ where either $\phi_0' \approx_\alpha \phi_1'$ and $\phi_0'' = \phi_1''$, or vice-versa; in either case, we get $\phi_0' \equiv_\alpha \phi_1'$ and $\phi_0'' \equiv_\alpha \phi_1''$. Now apply similar reasoning to $\phi_1 \approx_\alpha \phi_2$, $\phi_2 \approx_\alpha \phi_3$, etc., to eventually get that $\psi = \psi' \wedge \psi''$ with $\phi' \equiv_\alpha \psi'$ and $\phi'' \equiv_\alpha \psi''$. Thus

$$\begin{aligned}\phi[\sigma] &= \phi'[\sigma] \wedge \phi''[\sigma] \\ &\equiv_\alpha \psi'[\sigma] \wedge \psi''[\sigma] \quad \text{by IH and Proposition 3.17} \\ &= \psi[\sigma].\end{aligned}$$

- The cases $\vee$ and $\neg$ are similar.
- Finally, suppose $\phi = \exists x\,\phi'$. There are two possibilities for $\phi = \phi_0 \approx_\alpha \phi_1$: either

$$\phi = \exists x\,\phi' \sim_\alpha \exists y\,\phi'[x \mapsto y] = \phi_1 \quad \text{with } y \notin \mathrm{FV}(\phi') \cup \{x\} \text{ and } \phi'[x \mapsto y] \text{ safe},$$

or

$$\phi = \exists x\,\phi' \approx_\alpha \exists x\,\phi_1' = \phi_1 \quad \text{with } \phi' \approx_\alpha \phi_1';$$

in both cases, call $\exists x_1 \phi_1' := \phi_1$. Similarly breaking down $\phi_1 \approx_\alpha \phi_2$, $\phi_2 \approx_\alpha \phi_3$, etc., we get

$$\phi = \exists x\,\phi' =: \underbrace{\exists x_0\,\phi_0'}_{\phi_0} \approx_\alpha \underbrace{\exists x_1\,\phi_1'}_{\phi_1} \approx_\alpha \underbrace{\exists x_2\,\phi_2'}_{\phi_2} \approx_\alpha \cdots \approx_\alpha \underbrace{\exists x_n\,\phi_n'}_{\phi_n} := \exists y\,\psi' = \psi$$

where each $\approx_\alpha$ is either because of $\sim_\alpha$ (in which case the variables are different), or because the inner formulas satisfy $\approx_\alpha$ (in which case the variables are the same).

Let z be any variable which is not any of the x_i's, and does not occur free or bound in any of the ϕ_i''s; thus by Exercise 3.22(b), each $\phi_i'[x_i \mapsto z]$ is safe. For each of the above $\approx_\alpha$'s, say $\exists x_i\,\phi_i' \approx_\alpha \exists x_{i+1}\,\phi_{i+1}'$, we claim that $\phi_i'[x_i \mapsto z] \equiv_\alpha \phi_{i+1}'[x_{i+1} \mapsto z]$:

– If $\exists x_i\, \phi'_i \sim_\alpha \exists x_{i+1}\, \phi'_{i+1}$, with $x_{i+1} \notin \mathrm{FV}(\phi'_i) \cup \{x_i\}$ and $\phi'_{i+1} = \phi'_i[x_i \mapsto x_{i+1}]$, then

$$\begin{aligned}\phi'_{i+1}[x_{i+1} \mapsto z] &= \phi'_i[x_i \mapsto x_{i+1}][x_{i+1} \mapsto z] \\ &= \phi'_i[x_i \mapsto z]\end{aligned}$$

using Lemma 3.8 (where $(x_i \mapsto x_{i+1})[x_{i+1} \mapsto z] = (x_i \mapsto z)$, since $x_{i+1} \notin \mathrm{FV}(\phi'_i)$).

– Otherwise, $\exists x_i\, \phi'_i \approx_\alpha \exists x_{i+1}\, \phi'_{i+1}$ holds because $x_i = x_{i+1}$ and $\phi'_i \approx_\alpha \phi'_{i+1}$. Then

$$\phi'_i[x_i \mapsto z] \equiv_\alpha \phi'_{i+1}[x_{i+1} \mapsto z]$$

since these substitutions are safe, and the IH gives us Proposition 3.20 for $\phi'_i \equiv_\alpha \phi'_{i+1}$.

We have shown

$$\phi'[x \mapsto z] = \phi'_0[x_0 \mapsto z] \equiv_\alpha \phi'_1[x_1 \mapsto z] \equiv_\alpha \cdots \equiv_\alpha \phi'_n[x_n \mapsto z] = \psi'[y \mapsto z]$$

for all but finitely many z, which proves the weaker version of Proposition 3.28(e).

To complete the induction, we need to prove Proposition 3.20 for $\phi = \exists x\, \phi' \equiv_\alpha \exists y\, \psi' = \psi$. By restricting σ, we may assume $X = \mathrm{FV}(\exists x\, \phi') = \mathrm{FV}(\exists y\, \psi')$. Let z be one of the all-but-finitely-many variables as above, which is also not in either X or any term in the image of σ. Then since $\phi'[x \mapsto z] \equiv_\alpha \psi'[y \mapsto z]$ as shown above, and $\mathrm{HT}(\phi'[x \mapsto z]) = \mathrm{HT}(\phi') < \mathrm{HT}(\phi)$ (by Exercise 3.29), we may apply the IH to get

$$\phi'[x \mapsto z][\sigma] \equiv_\alpha \psi'[y \mapsto z][\sigma],$$

whence

$$\begin{aligned}(\exists x\, \phi')[\sigma] &= \exists x\, \phi'[\sigma\langle x \mapsto x\rangle] \\ &\sim_\alpha \exists z\, \phi'[\sigma\langle x \mapsto x\rangle][x \mapsto z] && \text{since } z \text{ does not appear free or bound in } \phi \\ &= \exists z\, \phi'[\sigma\langle x \mapsto z\rangle] && \text{by Lemma 3.8, since } z \text{ does not appear in } \mathrm{im}(\sigma) \\ &= \exists z\, \phi'[x \mapsto z][\sigma] && \text{by Lemma 3.8, since } z \notin X \\ &\equiv_\alpha \exists z\, \psi'[y \mapsto z][\sigma] \\ &\sim_\alpha (\exists y\, \psi')[\sigma] && \text{similarly.}\end{aligned}$$

We have now proved Proposition 3.20, as well as Proposition 3.28 with the weaker version of (e). To prove the original (e), where z is any variable such that $\phi'[x \mapsto z]$ and $\psi'[y \mapsto z]$ are both safe: by the weaker version, we may find some other variable $z' \notin \mathrm{FV}(\phi') \cup \mathrm{FV}(\psi')$ such that

$$\phi'[x \mapsto z'] \equiv_\alpha \psi'[y \mapsto z']$$

and these substitutions are both safe; now apply Proposition 3.20 and Lemma 3.8 to get

$$\phi'[x \mapsto z] = \phi'[x \mapsto z'][z' \mapsto z] \equiv_\alpha \psi'[y \mapsto z'][z' \mapsto z] = \psi'[y \mapsto z]. \qquad \square$$

One application of Proposition 3.28 is to rigorously prove that certain formulas *aren't* α-equivalent, instead of arguing informally about relative "positions" of free versus bound variables:

Example 3.30. We have

$$\begin{aligned}&\forall x\, (0 \le x \to \exists y\, (y \cdot y = x)) \equiv_\alpha \forall z\, (0 \le z \to \exists z\, (z \cdot z = z)) \\ \iff\quad &0 \le w \to \exists y\, (y \cdot y = w) \equiv_\alpha 0 \le w \to \exists z\, (z \cdot z = z);\end{aligned}$$

$\implies$ is by Proposition 3.28(e) using the variable w, since both $(0 \le x \to \exists y\, (y \cdot y = x))[x \mapsto w]$ and $(0 \le z \to \exists z\, (z \cdot z = z))[z \mapsto w]$ are safe (while $\impliedby$ is because we can use Proposition 3.17 to add a $\forall w$ to each side, and then change w back to x on the LHS and z on the RHS). This is in turn

$$\iff\quad 0 \le w \equiv_\alpha 0 \le w \text{ and } \exists y\, (y \cdot y = w) \equiv_\alpha \exists z\, (z \cdot z = z),$$

which is false, because for example the free variables of the last two formulas are different (using Proposition 3.18).

4. First-order proofs

4.1. **Natural deduction for first-order logic.** We now define a natural deduction system for first-order logic. As in propositional logic, we will design the system so as to capture the informal proofs that mathematicians write in practice. Here is an example of an informal first-order proof:

Example 4.1. In every abelian group, for every x, y, there is a z such that $x + z = y$. $(\mathcal{T}_{\mathsf{abgrp}} \vdash \forall x\, \forall y\, \exists z\, (x+z=y))$

Proof. Let x, y be arbitrary; we must find z such that $x + z = y$. $(\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} \exists z\, (x+z=y))$

We have

$$\left.\begin{aligned} x + ((-x) + y) &= (x + (-x)) + y && \text{by associativity} \\ &= 0 + y && \text{by inverse law} \\ &= y + 0 && \text{by commutativity} \\ &= y && \text{by identity law.} \end{aligned}\right\} (\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x+((-x)+y)=y)$$

Thus $z := (-x) + y$ works.

So indeed, for every x, y, there is a z such that $x + z = y$. □

Compared to propositional proofs, we see that, at each intermediate stage of the above proof, not only have we made some background assumptions (which in the above proof are always just the abelian group axioms, $\mathcal{T}_{\mathsf{abgrp}}$), but we may also have fixed some free variables (x, y above). Furthermore, while this does not happen in the above example, in general the free variables may need to be mentioned in the background assumptions as well; recall that our definition of *theory* in Section 2.3 does not allow free variables.

Definition 4.2. Let $\mathcal{A}$ be a first-order signature. By an **open $\mathcal{A}$-theory** $\mathcal{T}$, we mean a set of arbitrary first-order $\mathcal{A}$-formulas, possibly with free variables; we sometimes call an $\mathcal{A}$-theory in the original sense of Section 2.3 a **closed $\mathcal{A}$-theory** for emphasis. If $\mathcal{T} \subseteq \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$, i.e., all the free variables in $\mathcal{T}$ are from X, then we call $\mathcal{T}$ a **theory with free variables from** X.

A **first-order $\mathcal{A}$-sequent** is an expression of the form

$$\mathcal{T} \vdash_X \phi,$$

read "$\mathcal{T}$ proves ϕ under X", where $X \subseteq \mathcal{V}$ is a set of variables, $\mathcal{T} \subseteq \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$ is an open $\mathcal{A}$-theory with free variables from X, and ϕ is an $\mathcal{A}$-formula with free variables from X. Informally, this denotes the assertion ϕ under the background assumptions $\mathcal{T}$ and the fixed variables X.

In Example 4.1, we have labelled the sequents in three of the subproofs. We have not yet labelled any of the inference rules used to go between these sequents, because they all concern the new features of first-order formulas: quantifiers $\forall, \exists$ and equality $=$. (See Definition 4.5 below.)

However, the vast majority of the inference rules in first-order logic are actually the same as in propositional logic: the informal principles of reasoning that those rules capture are also valid when reasoning *about elements*. The following definitions allow us to import "for free" things we have already seen in propositional logic:

Definition 4.3. Let $\mathcal{A}$ be a (propositional) alphabet, $\mathcal{B}$ be a first-order signature. A **formula substitution from $\mathcal{A}$ to $\mathcal{B}$ with free variables from** X is a function $\sigma : \mathcal{A} \to \mathcal{L}^X_{\mathrm{form}}(\mathcal{B})$. Given such a σ, we define its **substitution into** a propositional $\mathcal{A}$-formula $\phi \in \mathcal{L}(A)$, resulting in a first-order $\mathcal{B}$-formula $\phi[\sigma]$ exactly as on HW2:

$$\begin{aligned} P[\sigma] &:= \sigma(P) \quad \text{for } P \in \mathcal{A}, & (\neg\phi)[\sigma] &:= \neg(\phi[\sigma]), \\ (\phi \wedge \psi)[\sigma] &:= \phi[\sigma] \wedge \psi[\sigma], & \top[\sigma] &:= \top, \\ (\phi \vee \psi)[\sigma] &:= \phi[\sigma] \vee \psi[\sigma], & \bot[\sigma] &:= \bot. \end{aligned}$$

One can easily prove the following analog to Lemma 3.7 (but much easier, since propositional formulas do not bind variables):

Exercise 4.4. For $\phi \in \mathcal{L}(\mathcal{A})$ and $\sigma : \mathcal{A} \to \mathcal{L}^X_{\text{form}}(\mathcal{B})$, we have $\text{FV}(\phi[\sigma]) = \bigcup_{P \in \text{AT}(\phi)} \text{FV}(\sigma(P))$, where $\text{AT}(\phi)$ is the set of atomic formulas appearing in ϕ, as in Example 1.6 in the notes on propositional logic. In particular, $\text{FV}(\phi[\sigma]) \subseteq X$, i.e., $\phi[\sigma] \in \mathcal{L}^X_{\text{form}}(\mathcal{B})$.

We are now ready to give

Definition 4.5. The **natural deduction system for first-order logic**, over the set of first-order sequents, has the following inference rule(schema)s:

- For every propositional inference rule
$$\frac{\mathcal{T}_1 \vdash \phi_1 \quad \cdots \quad \mathcal{T}_n \vdash \phi_n}{\mathcal{T} \vdash \phi},$$
every **first-order instance** of it, obtained by performing the *same* formula substitution $\sigma : \mathcal{B} \to \mathcal{L}^X_{\text{form}}(\mathcal{A})$ into all the formulas in it, resulting in
$$\frac{\mathcal{T}_1[\sigma] \vdash_X \phi_1[\sigma] \quad \cdots \quad \mathcal{T}_n[\sigma] \vdash_X \phi_n[\sigma]}{\mathcal{T}[\sigma] \vdash_X \phi[\sigma]},$$
is declared to be a first-order inference rule. (Here, $\mathcal{T}[\sigma]$ of course means $\{\psi[\sigma] \mid \psi \in \mathcal{T}\}$.)
For example,
$$(\vee\text{I1}) \ \frac{\mathcal{T} \vdash_X \phi}{\mathcal{T} \vdash_X \phi \vee \psi} \quad \text{for } \mathcal{T} \subseteq \mathcal{L}^X_{\text{form}}(\mathcal{A}),\ \phi, \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$$
is a first-order instance of the propositional $(\vee\text{I1})$ rule. *This is not completely obvious from the definition:* after all, the formulas ϕ, ψ, as well as the formulas in $\mathcal{T}$, may contain quantifiers. To see this in a systematic manner, for each $\theta \in \mathcal{T}$, let R_θ be an atomic propositional formula; then the above $(\vee\text{I1})$ is a first-order instance of the propositional rule
$$(\vee\text{I1}) \ \frac{\{R_\theta \mid \theta \in \mathcal{T}\} \vdash P}{\{R_\theta \mid \theta \in \mathcal{T}\} \vdash P \vee Q}$$
via the substitution $\sigma : \{R_\theta \mid \theta \in \mathcal{T}\} \cup \{P, Q\} \to \mathcal{L}^X_{\text{form}}(\mathcal{A})$ mapping $R_\theta \mapsto \theta$, $P \mapsto \phi$, $Q \mapsto \psi$.
- For $=$, we have the following rules:
$$(=\text{I}) \ \frac{}{\mathcal{T} \vdash_X t = t} \quad \text{for } t \in \mathcal{L}^X_{\text{term}}(\mathcal{A}),$$
$$(=\text{E}) \ \frac{\mathcal{T} \vdash_X s = t \quad \mathcal{T} \vdash_X \phi[x \mapsto s]}{\mathcal{T} \vdash_X \phi[x \mapsto t]} \quad \text{for } s, t \in \mathcal{L}^X_{\text{term}}(\mathcal{A}),\ \phi \in \mathcal{L}^{X \cup \{x\}}_{\text{form}}(\mathcal{A}) \text{ such that } \phi[x \mapsto s], \phi[x \mapsto t] \text{ are safe.}$$
The $(=\text{I})$ rule says that everything equals itself, while the $(=\text{E})$ rule (sometimes called the **Leibniz rule**) says that things which are equal are interchangeable in every statement; we call ϕ here the **template formula** in which the two equal things can be swapped.
- Finally, for $\exists$, we have the following rules:
$$(\exists\text{I}) \ \frac{\mathcal{T} \vdash_X \phi[x \mapsto t]}{\mathcal{T} \vdash_X \exists x\, \phi} \quad \text{for } \phi \in \mathcal{L}^{X \cup \{x\}}_{\text{form}}(\mathcal{A}),\ t \in \mathcal{L}^X_{\text{term}}(\mathcal{A}) \text{ such that } \phi[x \mapsto t] \text{ is safe,}$$
$$(\exists\text{E}) \ \frac{\mathcal{T} \vdash_X \exists x\, \phi \quad \mathcal{T} \cup \{\phi\} \vdash_{X \cup \{x\}} \psi}{\mathcal{T} \vdash_X \psi} \quad \text{for } \phi \in \mathcal{L}^{X \cup \{x\}}_{\text{form}}(\mathcal{A}),\ \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A}) \text{ such that } x \notin X.$$
The $(\exists\text{I})$ rule says "to prove $\exists x\, \phi$, produce a witness t", while the $(\exists\text{E})$ rule says "to use $\exists x\, \phi$ to prove ψ, fix x such that ϕ, and prove ψ"; the restriction $x \notin X$ says that x is a *newly* fixed variable, about which the only thing we know is ϕ (since $\text{FV}(\mathcal{T}) \subseteq X$).

As in propositional logic, if there is a deduction of the *sequent* $\mathcal{T} \vdash_X \phi$ from *no* hypotheses, then we also write

$$\mathcal{T} \vdash_X \phi$$

as a ("meta") *statement*, and say $\mathcal{T}$ **proves** ϕ **(under** X**)**, or that ϕ is a **provable consequence of** $\mathcal{T}$ **(under** X**)**. We allow ourselves to omit $\mathcal{T}$ or X when it's empty; a similar edge-case ambiguity as in Remark 2.19 occurs here when X could be empty or not (see HW10).

The definitions of **derivable** and **admissible** inference rule are the same as in propositional logic: the former means there is a deduction of the given rule using only the basic rules above, while the latter means that deductions of the hypotheses of the given rule from *no* hypotheses may be transformed into a deduction of the conclusion from *no* hypotheses. A rule without hypotheses is derivable iff it is admissible, iff its conclusion is provable in the above sense. In general, every derivable rule is admissible, but not vice-versa, for the exact same reasons as in propositional logic.

4.2. **Examples of deductions and derivable/admissible rules.**

Example 4.6. *Every first-order instance of every provable sequent in propositional logic is provable.* The proof of this is exactly the same as on HW2: just perform the same formula substitution σ into the entire propositional deduction $\mathcal{D}$. Since we took every first-order instance of a propositional inference rule to be a first-order inference rule, the resulting first-order deduction $\mathcal{D}[\sigma]$ will still be a valid deduction.

Exercise 4.7. Verify this.

For example, recall that for any propositional formula ϕ, it and its double negation $\neg\neg\phi$ provably imply each other (see notes on propositional logic, Example 3.8). Here was one of the deductions:

$$\dfrac{\dfrac{\dfrac{}{\{\phi, \neg\phi\} \vdash \phi}(\mathrm{A}) \qquad \dfrac{}{\{\phi, \neg\phi\} \vdash \neg\phi}(\mathrm{A})}{\{\phi, \neg\phi\} \vdash \bot}(\neg\mathrm{E})}{\{\phi\} \vdash \neg\neg\phi}(\neg\mathrm{I})$$

We claim that the same holds for any *first-order* $\phi \in \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$ (over the variables X). Again, this is perhaps not as obvious as it looks; we need to first take ϕ above to be an atomic formula P, and *then* perform the substitution $P \mapsto \phi$ for the desired first-order ϕ, in order to arrive at the deduction

$$\dfrac{\dfrac{\dfrac{}{\{\phi, \neg\phi\} \vdash_X \phi}(\mathrm{A}) \qquad \dfrac{}{\{\phi, \neg\phi\} \vdash_X \neg\phi}(\mathrm{A})}{\{\phi, \neg\phi\} \vdash_X \bot}(\neg\mathrm{E})}{\{\phi\} \vdash_X \neg\neg\phi}(\neg\mathrm{I})$$

Example 4.8. Similarly, *every first-order instance of a derivable propositional rule is derivable.* The proof of this fact is exactly the same as for the no-hypothesis case (**Exercise**). For example, the following rules are derivable (by Example 3.17 in propositional logic and HW3):

$$(\mathrm{P})\ \dfrac{\mathcal{T} \cup \{\neg\phi\} \vdash_X \phi}{\mathcal{T} \vdash_X \phi} \qquad (\mathrm{LEM})\ \dfrac{}{\mathcal{T} \vdash_X \phi \vee \neg\phi} \qquad (\to\mathrm{I})\ \dfrac{\mathcal{T} \cup \{\phi\} \vdash_X \psi}{\mathcal{T} \vdash_X \phi \to \psi}$$

On the other hand, *it is not true that every first-order instance of an admissible propositional rule is automatically admissible!* Think about what this would mean: we would need to know that if all the hypotheses of the instance are provable, then so is the conclusion. *If* we knew that the proofs of the hypotheses of the instance were instances of proofs of the hypotheses of the original rule, *then* by admissibility of the original rule, its conclusion is provable, *hence* the conclusion of the instance rule is also provable. But a moment's thought reveals this to be an unreasonable "if":

Example 4.9. The propositional rule

$$\frac{\vdash P}{\vdash \bot}$$

is vacuously admissible: there is no proof of $\vdash P$ (by soundness, since P is not true under every truth assignment m). But the substitution $P \mapsto \top$ takes this to the (in fact propositional) instance

$$\frac{\vdash \top}{\vdash \bot}$$

which is no longer admissible, since there is a proof of $\vdash \top$ (by ($\top$I), which is *not* an instance of any proof of $\vdash P$), but there is still no proof of $\vdash \bot$ (by soundness).

Example 4.10. The propositional rule

$$(\wedge\text{I})\ \frac{\{P, Q \wedge R\} \vdash R \qquad \{P, Q \wedge R\} \vdash S}{\{P, Q \wedge R\} \vdash R \wedge S}$$

has the first-order instance

$$(\wedge\text{I})\ \frac{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} x \cdot y = 0 \qquad \{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} y \cdot z = 1}{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} (x \cdot y = 0) \wedge (y \cdot z = 1)}$$

via the substitution $\sigma : \{P, Q, R, S\} \to \mathcal{L}_{\text{form}}^{\{x,y,z\}}(\mathcal{A}_{\mathsf{ordfield}})$ mapping $Q \mapsto 0 \leq 1$, $R \mapsto x \cdot y = 0$ (which are the only possibilities for $\sigma(Q), \sigma(R)$, since $Q \wedge R$ needs to be mapped to a conjunction on the LHS), $S \mapsto y \cdot z = 1$ (by considering the RHS of the second hypothesis), and $P \mapsto (0 \leq 1) \wedge (x \cdot y = 0)$ (since P also needs to be mapped to some formula in the LHS theory).

On the other hand, the first-order rule

$$(\wedge\text{I})\ \frac{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} 0 \leq 1 \qquad \{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} y \cdot z = 1}{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} (0 \leq 1) \wedge (y \cdot z = 1)}$$

would *not* be a first-order instance of the above propositional rule, since R would need to be mapped to both $0 \leq 1$ and $x \cdot y = 0$ (and these are not the same, even up to α-equivalence; see Convention 4.21 below). Nonetheless, this rule *is* a first-order instance of a valid propositional ($\wedge$I), just not the one above; we just need to choose the propositional formulas in a more general fashion, e.g., by systematically assigning a different propositional symbol $P, Q, R, \ldots$ to each first-order formula, as explained in Definition 4.5:

$$(\wedge\text{I})\ \frac{\{P\} \vdash Q \qquad \{P\} \vdash R}{\{P\} \vdash Q \wedge R}$$

which becomes the above rule under $P \mapsto (0 \leq 1) \wedge (x \cdot y = 0)$, $Q \mapsto 0 \leq 1$, and $R \mapsto y \cdot z = 1$.

Finally, neither of

$$(\wedge\text{I})\ \frac{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y\}} 0 \leq 1 \qquad \{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y\}} y \cdot z = 1}{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y\}} (0 \leq 1) \wedge (y \cdot z = 1)},$$

$$(\wedge\text{I})\ \frac{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y\}} 0 \leq 1 \qquad \{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} y \cdot z = 1}{\{(0 \leq 1) \wedge (x \cdot y = 0)\} \vdash_{\{x,y,z\}} (0 \leq 1) \wedge (y \cdot z = 1)}$$

is a valid first-order instance of any propositional rule. In the first rule, the second hypothesis as well as conclusion are not valid first-order sequents, since the fixed variable set $\{x, y\}$ does not include all the free variables appearing in the formulas on either side. The second rule does not have this problem; however, all the variable sets appearing in a first-order instance must be the same (and all the substitutions must be obtained via the same substitution σ).

We now turn to examples of the new first-order rules. The (=I) rule is self-explanatory. Here is an example of (=E); pay close attention to the role of the template formula:

Example 4.11. Here is part of the proof that $0 \leq 1$ from the ordered field axioms (Example 2.25):

$$(=\text{E})\ \frac{\begin{array}{cc}\vdots & \vdots \\ \mathcal{T}_{\mathsf{ordfield}} \vdash 1 \cdot 1 = 1 & \mathcal{T}_{\mathsf{ordfield}} \vdash 0 \leq 1 \cdot 1\end{array}}{\mathcal{T}_{\mathsf{ordfield}} \vdash 0 \leq 1}$$

The template formula here is $\phi := (0 \leq x)$, and we are plugging $x \mapsto 1 \cdot 1$ and $x \mapsto 1$ into it.

(The rest of the proof will have to wait until we have the ($\forall$E) rule, so that we can make use of the $\forall$ axioms in $\mathcal{T}_{\mathsf{ordfield}}$; see Exercise 4.32.)

Example 4.12. Here is another (more artificial) application of (=E):

$$(=\text{E})\ \frac{(\text{A})\ \dfrac{}{\{x = -y,\ x \leq x\} \vdash_{\{x,y\}} x = -y} \qquad (\text{A})\ \dfrac{}{\{x = -y,\ x \leq x\} \vdash_{\{x,y\}} x \leq x}}{\{x = -y,\ x \leq x\} \vdash_{\{x,y\}} -y \leq -y}$$

Here, the template formula *could* be $\phi := (x \leq x)$, in which case we are substituting $x \mapsto x$ and $x \mapsto -y$ into it; but it is probably clearer to choose the template formula $\phi := (z \leq z)$ instead, to emphasize that z is the "hole" in which we are replacing the equal terms $x = -y$. If we wanted to replace only *one* of the x's, we have no choice but to use a different variable in the template:

$$(=\text{E})\ \frac{(\text{A})\ \dfrac{}{\{x = -y,\ x \leq x\} \vdash_{\{x,y\}} x = -y} \qquad (\text{A})\ \dfrac{}{\{x = -y,\ x \leq x\} \vdash_{\{x,y\}} x \leq x}}{\{x = -y,\ x \leq x\} \vdash_{\{x,y\}} -y \leq x}$$

Here, the template formula is $\phi := (z \leq x)$, with the substitutions $z \mapsto x$ and $z \mapsto -y$.

It is common to use (=E) indirectly, via one of the following familiar ways of reasoning about equality which are derived from (=E):

Example 4.13. The following **symmetry** rule for $=$

$$(\textsc{Sym})\ \frac{\mathcal{T} \vdash_X s = t}{\mathcal{T} \vdash_X t = s}$$

is derivable:

$$(=\text{E})\ \frac{\mathcal{T} \vdash_X s = t \qquad (=\text{I})\ \dfrac{}{\mathcal{T} \vdash_X s = s}}{\mathcal{T} \vdash_X t = s}$$

(Here we are applying (=E) to the template formula $x = s$ with substitutions $x \mapsto s, t$.)

Using this and ($\to$I) (from Example 4.8), we can give a deduction of $\vdash_X (s = t) \to (t = s)$:

$$(\to\text{I})\ \frac{(\text{A})\ \dfrac{}{\{s = t\} \vdash_X s = t}}{\{s = t\} \vdash_X t = s}}{\vdash_X (s = t) \to (t = s)}$$

Taking $s = x$, $t = y$, $X = \{x, y\}$, and using ($\forall$I) from Example 4.27 below, we can turn this into

$$\vdash \forall x\, \forall y\, ((x = y) \to (y = x)),$$

which is perhaps the most direct way of expressing "equality is symmetric".

Exercise 4.14 (HW9)**.** The following **transitivity** rule for $=$ is derivable:

$$(\textsc{Trans})\ \frac{\mathcal{T} \vdash_X r = s \qquad \mathcal{T} \vdash_X s = t}{\mathcal{T} \vdash_X r = t}$$

Example 4.15. The following **compatibility** or **congruence** rule for $=$ is derivable:

$$(\textsc{Cong})\ \frac{\mathcal{T} \vdash_X s_1 = t_1 \quad \cdots \quad \mathcal{T} \vdash_X s_n = t_n}{\mathcal{T} \vdash_X f(s_1, \ldots, s_n) = f(t_1, \ldots, t_n)} \quad \text{for } f \in \mathcal{A}^n_{\text{fun}}.$$

To prove this, we repeatedly apply (=E) (you should read this top-down):

$$(=\text{E})\ \frac{\mathcal{T} \vdash_X s_n = t_n \qquad \vdots}{\mathcal{T} \vdash_X f(s_1, \ldots, s_n) = f(t_1, \ldots, t_n)}$$

$$(=\text{E})\ \frac{\mathcal{T} \vdash_X s_2 = t_2 \qquad (=\text{E})\ \dfrac{\mathcal{T} \vdash_X s_1 = t_1 \qquad (=\text{I})\ \dfrac{}{\mathcal{T} \vdash_X f(s_1, \ldots, s_n) = f(s_1, \ldots, s_n)}}{\mathcal{T} \vdash_X f(s_1, \ldots, s_n) = f(t_1, s_2, \ldots, s_n)}}{\mathcal{T} \vdash_X f(s_1, \ldots, s_n) = f(t_1, t_2, s_3, \ldots, s_n)}$$

Exercise 4.16. To which template ϕ and substitutions $x \mapsto s, t$ are we applying each (=E)?

Example 4.17. We can now formalize part of the informal proof from Example 4.1, namely the chain of $=$'s in the middle:

$$(\textsc{Trans})\ \frac{\mathcal{D}_1 \qquad (\textsc{Trans})\ \dfrac{(\textsc{Cong})\ \dfrac{\mathcal{D}_2 = \dfrac{\vdots}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x + (-x) = 0} \qquad (=\text{I})\ \dfrac{}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} y = y}}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} (x + (-x)) + y = 0 + y} \qquad (\textsc{Trans})\ \dfrac{\mathcal{D}_3 \qquad \mathcal{D}_4}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} 0 + y = y}}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} (x + (-x)) + y = y}}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x + ((-x) + y) = y}$$

The subproofs $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3, \mathcal{D}_4$ need (∀E) to make use of the axioms in $\mathcal{T}_{\mathsf{abgrp}}$; see Example 4.28 below.

The next step *after* the above chain of $=$'s is to apply (∃I):

$$(\exists\text{I})\ \frac{\dfrac{\vdots}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x + ((-x) + y) = y}}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} \exists z\, (x + z = y)}$$

Here the witness term t in Definition 4.5 is of course $(-x) + y$.

To complete the proof of Example 4.1, we need to apply (∀I); again see Example 4.28 below.

Remark 4.18. It is often helpful to think of $\exists$ as analogous to $\vee$: $\phi_1 \vee \phi_2$ means ϕ_i is true for some $i = 1, 2$, whereas $\exists x\, \phi$ means ϕ is true for some x in the underlying set (which may be infinite).

The (∃I) rule is then analogous to (∨I1) and (∨I2) ("to prove $\exists x\, \phi$, prove ϕ for some x", namely the witness term t, versus "to prove $\phi_1 \vee \phi_2$, prove ϕ_i for some i").

Similarly, the (∃E) rule is analogous to (∨E) ("to use $\exists x\, \phi$ to prove ψ, prove ψ assuming ϕ for an arbitrary x" versus "to use $\phi_1 \vee \phi_2$ to prove ψ, prove ψ assuming ϕ_i for each $i = 1, 2$"). This is illustrated by the following example:

Example 4.19. In order to prove that in a field, an element with a reciprocal cannot be zero:

$$\begin{array}{l}(\text{A})\\(\exists\text{E})\end{array}\ \frac{\mathcal{T}_{\mathsf{field}} \cup \{\exists y\, (x \cdot y = 1)\} \vdash_{\{x\}} \exists y\, (x \cdot y = 1) \qquad (\neg\text{I})\ \dfrac{\dfrac{\vdots}{\mathcal{T}_{\mathsf{field}} \cup \{\exists y\, (x \cdot y = 1),\ x \cdot y = 1,\ x = 0\} \vdash_{\{x,y\}} \bot}}{\mathcal{T}_{\mathsf{field}} \cup \{\exists y\, (x \cdot y = 1),\ x \cdot y = 1\} \vdash_{\{x,y\}} \neg(x = 0)}}{\mathcal{T}_{\mathsf{field}} \cup \{\exists y\, (x \cdot y = 1)\} \vdash_{\{x\}} \neg(x = 0)}$$

Note how this application of (∃E) satisfies the condition $y \notin \{x\}$ from Definition 4.5, guaranteeing that y is a newly fixed variable about which nothing else is known.

(Again, the rest of the proof will use the (∀E) rule; see Exercise 4.30.)

Example 4.20. Here is an *invalid* application of (∃E), showing what could go wrong when we forget to check the condition that the variable is new:

$$\dfrac{\dfrac{}{\{\exists x\,(x=1)\}\vdash_{\{x\}}\exists x\,(x=1)}\,(\mathrm{A}) \qquad \dfrac{}{\{\exists x\,(x=1),\,x=1\}\vdash_{\{x\}} x=1}\,(\mathrm{A})}{\{\exists x\,(x=1)\}\vdash_{\{x\}} x=1}\,(\exists\mathrm{E})$$

Indeed, the conclusion says that under the assumption $\exists x\,(x=1)$ (true in any structure with a constant 1), we should have $x=1$ for an arbitrary x (clearly not true in general).

What if we have a formula $\exists x\,\phi$ that we want to use to prove something else, but we've already fixed the variable x? In informal proofs, we would say "fix y such that $\phi[x\mapsto y]$ holds" instead, where y is a new variable that doesn't clash with anything. Formally, this means we want to replace the formula $\exists x\,\phi$ by an α-equivalent copy $\exists y\,\phi[x\mapsto y]$ first, *and then* apply (∃E).

Recall that in Propositions 3.17 to 3.20, we proved that α-equivalence is preserved by "every important logical notion" we'd seen up to that point. But deductions do *not* naturally respect α-equivalence, due to the variable condition in (∃E), as well as the various conditions on substitutions being safe (in the rules (=E) and (∃I) in Definition 4.5). We therefore have to mandate it by fiat:[10]

Convention 4.21. From now on, we allow α-equivalent formulas to be swapped at any location in a first-order sequent or deduction. That is, formally, a sequent $\mathcal{T}\vdash_X\phi$ no longer consists of a set of formulas $\mathcal{T}$ and a single formula ϕ; rather, ϕ is henceforth an *α-equivalence class* of formulas, while $\mathcal{T}$ is a set of α-equivalence classes of formulas.

However, in order to keep the notation readable, we will continue to write them as individual formulas (e.g., $\{\phi,\psi\}\vdash_X\theta$, rather than equivalence classes as in $\{[\phi],[\psi]\}\vdash_X[\theta]$).

Note that all of our definitions from Section 4.1 still make sense, using our previous preservation results from Propositions 3.17 to 3.20. For instance, the condition in Definition 4.2 of first-order sequent $\mathcal{T}\vdash_X\phi$, that the free variables of all formulas involved must belong to X, continues to make sense when we pass to α-equivalence classes because α-equivalent formulas have the same free variables by Proposition 3.18. Likewise, the safe substitutions in Definition 4.5 are well-defined on $\equiv_\alpha$-classes by Corollary 3.26. Hence, we can always apply an inference rule that we want to, without having to worry about clashing variables; if the variables clash, we may simply replace the formula involved with an α-equivalent copy.

Example 4.22. The correct way to apply (∃E) in Example 4.20 is to first change $\exists x\,(x=1)$:

$$\dfrac{\dfrac{}{\{\exists x\,(x=1)\}\vdash_{\{x\}}\exists x\,(x=1)\equiv_\alpha\exists y\,(y=1)}\,(\mathrm{A}) \qquad \dfrac{???}{\{\exists x\,(x=1),\,y=1\}\vdash_{\{x,y\}} x=1}}{\{\exists x\,(x=1)\}\vdash_{\{x\}} x=1}\,(\exists\mathrm{E})$$

We can no longer complete the proof, as expected.

Example 4.23. Let us show that for any $\phi,\psi\in\mathcal{L}_{\mathrm{form}}^{X\cup\{x\}}(\mathcal{A})$, we have the provable equivalence

$$\vdash_X(\exists x\,(\phi\vee\psi))\leftrightarrow(\exists x\,\phi)\vee(\exists x\,\psi).$$

After applying (∧I) and (→I) as usual, this amounts to proving

$$\{\exists x\,(\phi\vee\psi)\}\vdash_X(\exists x\,\phi)\vee(\exists x\,\psi),\qquad\qquad\{(\exists x\,\phi)\vee(\exists x\,\psi)\}\vdash_X\exists x\,(\phi\vee\psi).$$

To prove the first sequent, we would naturally want to

[10] Strictly speaking, this convention isn't necessary for the proof system to work: it is possible to *prove, in the proof system without built-in α-equivalence*, that α-equivalent formulas are logically equivalent. However, this gets quite messy and painful, and isn't really in the spirit of "α-equivalent means equivalent for all logical purposes".

"Fix x such that $\phi \vee \psi$ holds, and then split into the cases where ϕ or ψ holds, in each case proving the respective clause on the RHS."

The only snag is that x could already have been fixed, i.e., maybe $x \in X$. To circumvent this, pick some $y \notin X$ which is also not bound in any of the above formulas (here using Convention 3.23 to ensure that there is such a variable y), so that $(\phi \vee \psi)[x \mapsto y]$ is safe (by Exercise 3.22) and so $\exists x\,(\phi \vee \psi) \sim_\alpha \exists y\,(\phi \vee \psi)[x \mapsto y]$. We can then proceed to formalize the above proof sketch:

$$\dfrac{\dfrac{}{\{\exists x\,(\phi \vee \psi)\} \vdash_X \overbrace{\exists y\,(\phi \vee \psi)[x \mapsto y]}^{\exists x\,(\phi \vee \psi) \equiv_\alpha}}\,(\mathrm{A}) \qquad \dfrac{\dfrac{}{\mathcal{T} \vdash_{X \cup \{y\}} (\phi \vee \psi)[x \mapsto y]}\,(\mathrm{A}) \qquad \dfrac{\dfrac{\dfrac{}{\mathcal{T} \cup \{\phi[x \mapsto y]\} \vdash_{X \cup \{y\}} \phi[x \mapsto y]}\,(\mathrm{A})}{\mathcal{T} \cup \{\phi[x \mapsto y]\} \vdash_{X \cup \{y\}} \exists x\,\phi}\,(\exists \mathrm{I})}{\mathcal{T} \cup \{\phi[x \mapsto y]\} \vdash_{X \cup \{y\}} (\exists x\,\phi) \vee (\exists x\,\psi)}\,(\vee \mathrm{I1}) \quad \vdots}{\mathcal{T} := \{\exists x\,(\phi \vee \psi),\, (\phi \vee \psi)[x \mapsto y]\} \vdash_{X \cup \{y\}} (\exists x\,\phi) \vee (\exists x\,\psi)}\,(\vee \mathrm{E})}{\{\exists x\,(\phi \vee \psi)\} \vdash_X (\exists x\,\phi) \vee (\exists x\,\psi)}\,(\exists \mathrm{E})$$

where the last missing sub-deduction on the right is similar to the one to its left, and where in applying $(\exists \mathrm{I})$ we're using that $\phi[x \mapsto y]$ is safe since $(\phi \vee \psi)[x \mapsto y]$ is.

The deduction of the converse sequent $\{(\exists x\,\phi) \vee (\exists x\,\psi)\} \vdash_X \exists x\,(\phi \vee \psi)$ is similar in spirit, with the roles of $\exists$ and $\vee$ swapped, again illustrating Remark 4.18.

Exercise 4.24. Give the deduction of the converse, carefully stating your choice of new variable(s) and justifying uses of the inference rules where variable clashes might occur.

In order to give more interesting examples of first-order deductions, as well as to finish some of the examples from above, we now need to prove

Proposition 4.25 (weakening)**.** The following rule is admissible, for $\mathcal{T} \subseteq \mathcal{T}' \subseteq \mathcal{L}^X_{\mathrm{form}}(\mathcal{A})$:

$$(\mathrm{W})\ \frac{\mathcal{T} \vdash_X \phi}{\mathcal{T}' \vdash_X \phi}$$

(Note that for now, the variable set has to remain the same. In Corollary 4.35 below, we will see another version of weakening which allows us to introduce extra variables.)

Proof. By induction on the deduction of $\mathcal{T} \vdash_X \phi$, similarly to the propositional case (see Proposition 3.12 in notes); the point is that in all of the new first-order inference rules in Definition 4.5, we may also freely introduce extra assumptions into the theory. □

Example 4.26. It follows that *every first-order instance of an admissible propositional rule which was derived using weakening is now admissible.* This is because we may perform the formula substitution into the deduction of the propositional rule, as in Example 4.8 for derivable rules, and then replace all the resulting first-order instances of propositional (W) using the first-order (W) above. For example, we get that the following first-order rules are admissible:

$$(\to \mathrm{E})\ \frac{\mathcal{T} \vdash_X \phi \to \psi \qquad \mathcal{T} \vdash_X \phi}{\mathcal{T} \vdash_X \psi} \qquad\qquad (\textsc{Cut})\ \frac{\mathcal{T} \vdash_X \phi \qquad \mathcal{T} \cup \{\phi\} \vdash_X \psi}{\mathcal{T} \vdash_X \psi}$$

Example 4.27. Recalling that $\forall x$ is an abbreviation for $\neg \exists x \neg$, we have the following admissible rules for $\forall$ ("to prove $\forall x\,\phi$, let x be arbitrary and prove ϕ", and "from $\forall x\,\phi$, we may deduce ϕ for any particular x"):

$$(\forall \mathrm{I})\ \frac{\mathcal{T} \vdash_{X \cup \{x\}} \phi}{\mathcal{T} \vdash_X \forall x\,\phi} \quad \begin{array}{l}\text{for } \phi \in \mathcal{L}^{X \cup \{x\}}_{\mathrm{form}}(\mathcal{A}) \\ \text{such that } x \notin X,\end{array}$$

$$(\forall \mathrm{E})\ \frac{\mathcal{T} \vdash_X \forall x\,\phi}{\mathcal{T} \vdash_X \phi[x \mapsto t]} \quad \begin{array}{l}\text{for } \phi \in \mathcal{L}^{X \cup \{x\}}_{\mathrm{form}}(\mathcal{A}),\ t \in \mathcal{L}^X_{\mathrm{term}}(\mathcal{A}) \\ \text{such that } \phi[x \mapsto t] \text{ is safe.}\end{array}$$

Note the similarity to the rules for $\exists$ but with intro and elim swapped. As in Remark 4.18, it is helpful to think of these rules as analogous to those for $\wedge$: $\forall$ is like a conjunction indexed by the underlying set. Often more helpful, however, is a *different* analogy:[11] $\forall$ is analogous to $\to$ ("if ..., then ...", versus "if we have some x, then ..."). This latter analogy is especially evident in their respective inference rules: both sets of rules are derived/admissible; the ($\to$I) and ($\forall$I) rules are clear parallels ("assume ...", versus "assume x is arbitrary"); and we can also view ($\to$E) and ($\forall$E) as parallels ("prove ..., and deduce ...", versus "produce an x, and deduce ... for it").

The ($\forall$I) rule is derivable using (W), hence admissible:

$$\dfrac{\dfrac{\dfrac{}{\mathcal{T} \cup \{\exists x\, \neg\phi\} \vdash_X \exists x\, \neg\phi}\,(\mathrm{A}) \qquad \dfrac{\dfrac{\mathcal{T} \vdash_{X \cup \{x\}} \phi}{\mathcal{T}' \vdash_{X \cup \{x\}} \phi}\,(\mathrm{W}) \qquad \dfrac{}{\mathcal{T}' \vdash_{X \cup \{x\}} \neg\phi}\,(\mathrm{A})}{\mathcal{T}' := \mathcal{T} \cup \{\exists x\, \neg\phi, \neg\phi\} \vdash_{X \cup \{x\}} \bot}\,(\neg\mathrm{E})}{\mathcal{T} \cup \{\exists x\, \neg\phi\} \vdash_X \bot}\,(\exists\mathrm{E})}{\mathcal{T} \vdash_X \neg\exists x\, \neg\phi}\,(\neg\mathrm{I})$$

Note that the condition on x in ($\forall$I) implies the required condition in the application of ($\exists$E).

The ($\forall$E) rule is also derivable using (W), hence admissible:

$$\dfrac{\dfrac{\dfrac{}{\mathcal{T} \cup \{\neg\phi[x \mapsto t]\} \vdash_X \neg\phi[x \mapsto t]}\,(\mathrm{A})}{\mathcal{T} \cup \{\neg\phi[x \mapsto t]\} \vdash_X \exists x\, \neg\phi}\,(\exists\mathrm{I}) \qquad \dfrac{\mathcal{T} \vdash_X \neg\exists x\, \neg\phi}{\mathcal{T} \cup \{\neg\phi[x \mapsto t]\} \vdash_X \neg\exists x\, \neg\phi}\,(\mathrm{W})}{\mathcal{T} \cup \{\neg\phi[x \mapsto t]\} \vdash_X \bot}\,(\neg\mathrm{E})}{\mathcal{T} \vdash_X \phi[x \mapsto t]}\,(\mathrm{C})$$

Again, the conditions in ($\forall$E) imply the required conditions in ($\exists$I).

Example 4.28. To finish Example 4.17 (formalizing Example 4.1):

$$\dfrac{\dfrac{\mathcal{D}_1 \qquad \dfrac{\dfrac{\dfrac{}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} \forall x\, (x + (-x) = 0)}\,(\mathrm{A})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x + (-x) = 0}\,(\forall\mathrm{E}) \qquad \dfrac{}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} y = y}\,(=\mathrm{I})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} (x + (-x)) + y = 0 + y}\,(\textsc{Cong}) \qquad \dfrac{\dfrac{\dfrac{\dfrac{}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} \forall x\, \forall y\, (x + y = y + x)}\,(\mathrm{A})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} \forall y\, (0 + y = y + 0)}\,(\forall\mathrm{E})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} 0 + y = y + 0}\,(\forall\mathrm{E}) \qquad \mathcal{D}_4}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} 0 + y = y}\,(\textsc{Trans})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} (x + (-x)) + y = y}\,(\textsc{Trans})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x + ((-x) + y) = y}\,(\textsc{Trans})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x\}} \forall y\, (x + ((-x) + y) = y)}\,(\forall\mathrm{I})}{\mathcal{T}_{\mathsf{abgrp}} \vdash_{\varnothing} \forall x\, \forall y\, (x + ((-x) + y) = y)}\,(\forall\mathrm{I})$$

where the grey parts are as in Example 4.17, and the three applications of ($\forall$E) (reading from left to right, then top to bottom) are with the substitutions $x \mapsto x$, $x \mapsto 0$, and $y \mapsto y$, respectively.

Exercise 4.29. Fill in the sub-deductions $\mathcal{D}_1, \mathcal{D}_4$.

Exercise 4.30. In every commutative ring (Example 2.25), we have the following laws:

$$x \cdot 0 = 0,$$
$$x \cdot -y = -(x \cdot y).$$

The first can be thought of as a "0-ary" version of the distributive law, while the second is a "-1-ary" version. To prove these laws, note that the binary distributive law says that for each fixed x, the

[11] The counterpart of this other analogy, for $\exists$, is with $\wedge$ (instead of $\vee$): $\exists x\, \phi$ means "we have some x, *and* ϕ holds". These alternative analogies also play nicely with the fact that "$\exists x \in A\, (x \in B)$" means "$\exists x\, (x \in A \wedge x \in B)$", while "$\forall x \in A\, (x \in B)$" means "$\forall x\, (x \in A \to x \in B)$".

function $y \mapsto x \cdot y$ is a $\{+\}$-homomorphism, hence (since a commutative ring is an abelian group under addition) also preserves 0, $-$ by Proposition 2.36. One can also explicitly write out the arguments in Proposition 2.36 for these homomorphisms $y \mapsto x \cdot y$.

Formalize these arguments into deductions of

$$\mathcal{T}_{\mathsf{commring}} \vdash \forall x\,(x \cdot 0 = 0),$$
$$\mathcal{T}_{\mathsf{commring}} \vdash \forall x\,\forall y\,(x \cdot -y = -(x \cdot y)).$$

Using the first of these, and the field axiom $\neg(0 = 1)$, finish Example 4.19.

Exercise 4.31. Give a deduction of

$$\mathcal{T}_{\mathsf{commring}} \vdash \forall x\,(-(-x) = x).$$

[Informal proof: $-(-x) = 0 + (-(-x)) = x + (-x) + (-(-x)) = x + 0 = x$.]

Exercise 4.32. Formalize the following statement into a sequent, and give a deduction of it:

"In an ordered field, the square of every element is ≥ 0."

[Informal proof: since $\leq$ is a total order, every x is either ≤ 0 or ≥ 0. If $x \geq 0$, then since multiplication by nonnegative elements is order-preserving (by one of the ordered field axioms), we get $x \cdot x \geq 0 \cdot x = 0$. If $x \leq 0$, then adding $-x$ to both sides yields $0 \leq -x$, whence by the previous case, $0 \leq (-x)^2 = x^2$.]

Using this (and the identity axiom for $\cdot$), finish Example 4.11.

4.3. **Rules for variables.**

Proposition 4.33 (substitution)**.** The following rule is admissible, for any variable substitution $\sigma : X \to \mathcal{L}^Y_{\mathrm{term}}(\mathcal{A})$ such that $\mathcal{T}[\sigma], \phi[\sigma]$ are safe:

$$\text{(S)}\ \frac{\mathcal{T} \vdash_X \phi}{\mathcal{T}[\sigma] \vdash_Y \phi[\sigma]}$$

(where $\mathcal{T}[\sigma] := \{\psi[\sigma] \mid \psi \in \mathcal{T}\}$ is called safe if each $\psi[\vec{s}]$ is).

Example 4.34. From Example 4.17, we have $\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} \exists z\,(x + z = y)$. By (S) with $x \mapsto y$, we get $\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{y\}} \exists z\,(y + z = y)$. (Note that since $\mathcal{T}_{\mathsf{abgrp}}$ consists of sentences, $\mathcal{T}_{\mathsf{abgrp}}[\sigma] = \mathcal{T}_{\mathsf{abgrp}}$.)

On the other hand, if we had admissibility of (S) with $x \mapsto z$ (which is not safe for substitution into $\exists z\,(x + z = y)$), we would get $\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{y\}} \exists z\,(z + z = y)$, which would violate soundness (Proposition 4.40) since the abelian group $\mathbb{Z}$ with the variable assignment $y \mapsto 1$ fails to satisfy this formula. This example was our original motivation, in Example 3.4, for introducing safe substitution.

As a special case of (S), we can take $X \subseteq Y$, and take σ to be the identity function $X \to X \subseteq Y$, substitution of which is always safe, yielding

Corollary 4.35 (variable weakening)**.** The following rule is admissible, for $X \subseteq Y \subseteq \mathcal{V}$:

$$\text{(S)}\ \frac{\mathcal{T} \vdash_X \phi}{\mathcal{T} \vdash_Y \phi}$$

Intuitively, this says that if we can prove ϕ after fixing some variables, the same proof should still apply after fixing some more extraneous variables. This is analogous to the "ordinary" weakening rule (W) (Proposition 4.25), which says that we can add some extraneous assumptions. The combination of the two says that we can first expand the set of variables, and then the theory:

$$\begin{array}{l}\text{(S)}\ \dfrac{\mathcal{T} \vdash_X \phi}{\mathcal{T} \vdash_Y \phi} \\ \text{(W)}\ \dfrac{\mathcal{T} \vdash_Y \phi}{\mathcal{T}' \vdash_Y \phi}\end{array}$$

This has the following partial converse:

Proposition 4.36 (syntactic compactness)**.** If $\mathcal{T} \vdash_X \phi$, then there are finite $\mathcal{T}' \subseteq \mathcal{T}$ and $X' \subseteq X$ such that (X' contains all free variables occurring in $\mathcal{T}', \phi$, and) $\mathcal{T}' \vdash_{X'} \phi$.

Proof. First, we prove that keeping X fixed, we may shrink $\mathcal{T}$ down to a finite $\mathcal{T}' \subseteq \mathcal{T}$. This is exactly the same as in propositional logic (Proposition 3.21 in notes), by induction on the deduction of $\mathcal{T} \vdash_X \phi$, using that each step of the proof only uses at most one formula in $\mathcal{T}$.

It thus suffices to assume that $\mathcal{T}$ is already finite to begin with, and prove that we may shrink X down to a finite $X' \subseteq X$ containing all the free variables in $\mathcal{T}$. We proceed by induction on the deduction of $\mathcal{T} \vdash_X \phi$.

- If the deduction ends with (A), then $X' := \mathrm{FV}(\mathcal{T}) \cup \mathrm{FV}(\phi)$ works.
- If the deduction ends with, say,
$$(\vee\mathrm{E})\ \frac{\mathcal{T} \vdash_X \phi \vee \psi \qquad \mathcal{T} \cup \{\phi\} \vdash_X \theta \qquad \mathcal{T} \cup \{\psi\} \vdash_X \theta}{\mathcal{T} \vdash_X \theta},$$
then by the IH, there are finite $X_1, X_2, X_3 \subseteq X$ such that
$$\mathcal{T} \vdash_{X_1} \phi \vee \psi, \qquad \mathcal{T} \cup \{\phi\} \vdash_{X_2} \theta, \qquad \mathcal{T} \cup \{\psi\} \vdash_{X_3} \theta.$$
Let $X' := X_1 \cup X_2 \cup X_3$. Then by variable weakening, we may replace X_1, X_2, X_3 above with X', whence by ($\vee$E), $\mathcal{T} \vdash_{X'} \theta$.
- The rest of the first-order instances of propositional inference rules are similarly handled.
- If the deduction ends with $(=\mathrm{I})\ \dfrac{}{\mathcal{T} \vdash_X t = t}$, then $X' := \mathrm{FV}(\mathcal{T}) \cup \mathrm{FV}(t)$ works.
- If the deduction ends with
$$(=\mathrm{E})\ \frac{\mathcal{T} \vdash_X s = t \qquad \mathcal{T} \vdash_X \phi[x \mapsto s]}{\mathcal{T} \vdash_X \phi[x \mapsto t]}$$
where $s, t \in \mathcal{L}^X_{\mathrm{term}}(\mathcal{A})$ with $\phi[x \mapsto s], \phi[x \mapsto t]$ safe, then similarly to the ($\vee$E) case above, we may find a finite $X' \subseteq X$ such that
$$\mathcal{T} \vdash_{X'} s = t, \qquad \mathcal{T} \vdash_{X'} \phi[x \mapsto s];$$
in particular, for the first sequent to make sense, we must have $s, t \in \mathcal{L}^{X'}_{\mathrm{term}}(\mathcal{A})$, so that we may apply (=E) to get $\mathcal{T} \vdash_{X'} \phi[x \mapsto t]$.
- The ($\exists$I) case is similar (except that we should explicitly include all free variables in the witness term into X', to make sure we are still allowed to apply ($\exists$I)).
- Finally, if the deduction ends with
$$(\exists\mathrm{E})\ \frac{\mathcal{T} \vdash_X \exists x\, \phi \qquad \mathcal{T} \cup \{\phi\} \vdash_{X \cup \{x\}} \psi}{\mathcal{T} \vdash_X \psi}$$
with $x \notin X$, then by the IH, there are finite $X_1 \subseteq X$ and $X_2 \subseteq X \cup \{y\}$ such that
$$\mathcal{T} \vdash_{X_1} \exists x\, \phi, \qquad \mathcal{T} \cup \{\phi\} \vdash_{X_2} \psi.$$
Let $X' := X_1 \cup (X_2 \cap X)$ (cf. the proof of syntactic compactness, Proposition 3.21, from propositional logic). Then $X_2 \subseteq X' \cup \{y\}$ (because $X_2 \subseteq X \cup \{x\}$), so by variable weakening,
$$\mathcal{T} \vdash_{X'} \exists x\, \phi, \qquad \mathcal{T} \cup \{\phi\} \vdash_{X' \cup \{x\}} \psi.$$
Moreover, $x \notin X'$ since $x \notin X \supseteq X'$, so we may apply ($\exists$E) to deduce $\mathcal{T} \vdash_{X'} \psi$. □

In the rest of this subsection, we give the proof of Proposition 4.33. This proof is rather technical, but the basic idea is straightforward enough: we should be able to simply perform the variable substitution σ throughout the entire deduction $\mathcal{D}$ of $\mathcal{T} \vdash_X \phi$. This is analogous to how one performs a *formula* substitution into a *propositional* deduction, as on HW2 (see Example 4.6).

The added difficulties are because, when substituting into a *first-order* sequent, one expects to encounter issues with variable capture:

- Even though in Proposition 4.33, we assumed that the substitution of σ into the conclusion $\mathcal{T} \vdash_X \phi$ of $\mathcal{D}$ is safe, this does *not* ensure that the substitution into every formula in $\mathcal{D}$ is safe, since $\mathcal{D}$ could contain complicated intermediate formulas that don't appear anywhere in its conclusion (for example, in the hypotheses of ($\vee$E) or ($\exists$E)).

  Exercise 4.37. Give an example of this.

  However, this is not a real issue for us: by our Convention 4.21, formulas in deductions are really α-equivalence classes anyway; and we can always pick an α-equivalent formula for which the substitution of σ is safe, by Corollary 3.25.
- A more serious issue is that even if the substitution of σ into every *formula* in $\mathcal{D}$ is safe, we could still end up with an invalid deduction, *because the resulting deduction may violate the condition* $x \notin X$ *in* ($\exists$E). Indeed, this can happen even in the special case where σ is the identity, i.e., we are performing variable weakening. For example, in trying to weaken

$$(\exists\text{E})\ \frac{\begin{array}{cc}\vdots & \vdots\\ \cdots \vdash_{\{x\}} \exists y\,(x \cdot y = 1) & \cdots \cup \{x \cdot y = 1\} \vdash_{\{x,y\}} \neg(x = 0)\end{array}}{\cdots \vdash_{\{x\}} \neg(x = 0)}$$

  to the bigger set of variables $\{x, y\} \supseteq \{x\}$, we obtain the invalid

$$(\exists\text{E})\ \frac{\begin{array}{cc}\vdots & \vdots\\ \cdots \vdash_{\{x,y\}} \exists y\,(x \cdot y = 1) & \cdots \cup \{x \cdot y = 1\} \vdash_{\{x,y\}} \neg(x = 0)\end{array}}{\cdots \vdash_{\{x,y\}} \neg(x = 0)}.$$

To handle this second type of issue, we say that **the substitution of** $\sigma : X \to \mathcal{L}^Y_{\mathbf{term}}(\mathcal{A})$ **into** $\mathcal{D}$ **is safe** if in no application of ($\exists$E) in $\mathcal{D}$ is the added variable x in Y.

Lemma 4.38. If $\mathcal{T} \vdash_X \phi$ via a deduction into which substitution of $\sigma : X \to \mathcal{L}^Y_{\text{term}}(\mathcal{A})$ is safe, then $\mathcal{T}[\sigma] \vdash_Y \phi[\sigma]$.

Proof. By induction on the deduction $\mathcal{D}$ of $\mathcal{T} \vdash_X \phi$.

- The first-order instances of propositional rules are all straightforward: just substitute σ safely into all formulas in sight, possibly after replacing them with α-equivalent copies via Corollary 3.25.
- If $\mathcal{D} = (=\text{I})\ \dfrac{}{\mathcal{T} \vdash_X t = t}$, then $\mathcal{T}[\sigma] \vdash_Y (t = t)[\sigma] = (t[\sigma] = t[\sigma])$ again by ($=$I).
- Suppose $\mathcal{D}$ ends with

$$(=\text{E})\ \frac{\mathcal{T} \vdash_X s = t \qquad \mathcal{T} \vdash_X \phi[x \mapsto s]}{\mathcal{T} \vdash_X \phi[x \mapsto t]}$$

  where both substitutions are safe. After substituting σ into everything in sight, we get

$$(*)\qquad (=\text{E})\ \frac{\mathcal{T}[\sigma] \vdash_Y s[\sigma] = t[\sigma] \qquad \mathcal{T}[\sigma] \vdash_Y \phi[x \mapsto s][\sigma]}{\mathcal{T}[\sigma] \vdash_Y \phi[x \mapsto t][\sigma]}.$$

  In order for this to be a valid application of ($=$E), we need for $\phi[x \mapsto s][\sigma]$ and $\phi[x \mapsto t][\sigma]$ to be substitutions of $s[\sigma], t[\sigma]$ into some common template formula. By Lemma 3.8 (twice),

$$\phi[x \mapsto s][\sigma] = \phi[(x \mapsto s)[\sigma]] = \phi[\sigma\langle x \mapsto s[\sigma]\rangle] = \phi[\sigma\langle x \mapsto y\rangle][y \mapsto s[\sigma]]$$

where y is a *new* variable not in Y, hence not in any term in the image of σ, in order to ensure $\sigma\langle x \mapsto y\rangle[y \mapsto s[\sigma]] = \sigma\langle x \mapsto s[\sigma]\rangle$. Moreover, this last substitution is also safe, provided we first use Proposition 3.24 to replace ϕ with an α-equivalent formula none of whose bound variables appear free in Y (hence in $s[\sigma]$). Similarly,

$$\phi[x \mapsto t][\sigma] = \phi[(x \mapsto t)[\sigma]] = \phi[\sigma\langle x \mapsto t[\sigma]\rangle] = \phi[\sigma\langle x \mapsto y\rangle][y \mapsto t[\sigma]].$$

Thus $(*)$ is indeed a valid application of (=E), to the template formula $\phi[\sigma\langle x \mapsto y\rangle]$.

- Suppose $\mathcal{D}$ ends with

$$(\exists\text{I})\ \frac{\mathcal{T} \vdash_X \phi[x \mapsto t]}{\mathcal{T} \vdash_X \exists x\, \phi}$$

where $\phi[x \mapsto t]$ is safe. After substituting σ into everything in sight, we get

$$(\dagger) \qquad (\exists\text{I})\ \frac{\mathcal{T} \vdash_Y \phi[x \mapsto t][\sigma]}{\mathcal{T}[\sigma] \vdash_Y \exists x\, \phi[\sigma\langle x \mapsto x\rangle]}$$

Similarly to the (=E) case above, we have

$$\phi[x \mapsto t] = \phi[(x \mapsto t)[\sigma] = \phi[\sigma\langle x \mapsto t[\sigma]\rangle] = \phi[\sigma\langle x \mapsto y\rangle][y \mapsto t[\sigma]]$$

where y is chosen as above. Now in order for $(\dagger)$ to be a valid application of ($\exists$I), we note that the formula in its conclusion is

$$\begin{aligned}\exists x\, \phi[\sigma\langle x \mapsto x\rangle] &\sim_\alpha \exists y\, \phi[\sigma\langle x \mapsto x\rangle][x \mapsto y] \\ &= \exists y\, \phi[\sigma\langle x \mapsto y\rangle],\end{aligned}$$

using that the substitutions $\sigma\langle x \mapsto x\rangle[x \mapsto y]$ and $\sigma\langle x \mapsto y\rangle$ agree on all free variables in ϕ: they clearly agree on x; while for $z \in \mathrm{FV}(\phi) \setminus \{x\}$, we have $\sigma(z)[x \mapsto y] = \sigma(z)$ since $x \notin \mathrm{FV}(\sigma(z))$ by the original assumption (in the statement of (S)) that $(\exists x\, \phi)[\sigma]$ is safe.

- Finally, suppose $\mathcal{D}$ ends with

$$(\exists\text{E})\ \frac{\mathcal{T} \vdash_X \exists x\, \phi \qquad \mathcal{T} \cup \{\phi\} \vdash_{X \cup \{x\}} \psi}{\mathcal{T} \vdash_X \psi}$$

where $x \notin X$. This means $\sigma\langle x \mapsto x\rangle$ is simply σ extended by the identity function (without erasing any previous $\sigma(x)$). Thus, substituting σ into everything yields

$$(\exists\text{E})\ \frac{\mathcal{T}[\sigma] \vdash_Y \exists x\, \phi[\sigma] \qquad \mathcal{T}[\sigma] \cup \{\phi[\sigma]\} \vdash_{Y \cup \{x\}} \psi[\sigma]}{\mathcal{T}[\sigma] \vdash_Y \psi[\sigma]}$$

which is still a valid application of ($\exists$E) since $x \notin Y$ by our assumption that substitution of σ into $\mathcal{D}$ is safe. □

In order to finish proving Proposition 4.33, it thus remains to show that any deduction of $\mathcal{T} \vdash_X \phi$ can be turned into one into which substitution of σ is safe. The idea here is to think of the rule

$$(\exists\text{E})\ \frac{\begin{array}{c}\vdots \\ \mathcal{T} \vdash_X \exists x\, \phi\end{array} \qquad \begin{array}{c}\vdots \\ \mathcal{T} \cup \{\phi\} \vdash_{X \cup \{x\}} \psi\end{array}}{\mathcal{T} \vdash_X \psi}$$

as "binding" the free variable x in the second sub-deduction, much as a quantifier $\exists x\, \phi$ binds the free x in the subformula ϕ. When a substitution into this rule breaks the condition $x \notin X$, we should think of the ($\exists$E) as "capturing" the free variable x. The solution to variable capture is the familiar one: we need to replace the original deduction with an "α-equivalent" copy, where the free variable has been replaced with a new variable via a safe substitution. Since this is essentially similar to the arguments for α-equivalence of formulas in Section 3.2, the details are left to you:

Exercise 4.39.

(a) Prove by induction that if $\mathcal{T} \vdash_X \phi$, then for any infinite set Y disjoint from X, there is a deduction of $\mathcal{T} \vdash_X \phi$ whose new variables introduced by applications of ($\exists$E) all come from Y. [Imitate the proof of Proposition 3.24.]

(b) Conclude that for any $\sigma : X \to \mathcal{L}^Y_{\text{term}}(\mathcal{A})$, if $\mathcal{T} \vdash_X \phi$, then there is a deduction of it into which substitution of σ is safe. Thereby conclude Proposition 4.33.

4.4. **Soundness.** Let X be a set of variables, $\mathcal{T} \subseteq \mathcal{L}^X_{\text{form}}(\mathcal{A})$ be an open theory with free variables from X, and $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$. Extending the definition from Section 2.3 for closed theories, we write

$$\mathcal{T} \models_X \phi$$

if for all $\mathcal{A}$-structures $\mathcal{M}$ and $\alpha : X \to M$, if $\mathcal{M} \models_\alpha \psi$ for every $\psi \in \mathcal{T}$, then $\mathcal{M} \models_\alpha \phi$.

Proposition 4.40 (soundness)**.** If $\mathcal{T} \vdash_X \phi$, then $\mathcal{T} \models_X \phi$.

There is a subtlety hidden in the notation here: recall that by Convention 4.21, the sequent $\mathcal{T} \vdash_X \phi$ actually consists of α-equivalence classes of formulas. Thus, in order to even make sense of the claim $\mathcal{T} \models_X \phi$, we need to know that α-equivalent formulas always have the same interpretation; this is given by Lemma 4.42 below. In order to prove this, as well as soundness (since several inference rules refer to substitution), we will need to know how substitution is interpreted:

Lemma 4.41 (soundness of substitution, HW8)**.** Let $\sigma : X \to \mathcal{L}^Y_{\text{term}}(\mathcal{A})$ be a variable substitution, $\mathcal{M}$ be an $\mathcal{A}$-structure, and $\alpha : Y \to M$ be a variable assignment. We write

$$\begin{aligned} \sigma^{\mathcal{M}}(\alpha) : X &\longrightarrow M \\ x &\longmapsto \sigma(x)^{\mathcal{M}}(\alpha). \end{aligned}$$

(a) For a term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we have $t[\sigma]^{\mathcal{M}}(\alpha) = t^{\mathcal{M}}(\sigma^{\mathcal{M}}(\alpha))$.

(b) For a formula $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ such that $\phi[\sigma]$ is safe, we have $\phi[\sigma]^{\mathcal{M}}(\alpha) = \phi^{\mathcal{M}}(\sigma^{\mathcal{M}}(\alpha))$, i.e.,

$$\mathcal{M} \models_\alpha \phi[\sigma] \iff \mathcal{M} \models_{\sigma^{\mathcal{M}}(\alpha)} \phi.$$

Lemma 4.42 (soundness of α-equivalence)**.** If $\phi \equiv_\alpha \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, then $\models_X \phi \leftrightarrow \psi$.

Proof. Let $\mathcal{M}$ be an $\mathcal{A}$-structure and $\alpha : X \to M$ be a variable assignment. First, suppose $\exists x\, \phi \sim_\alpha \exists y\, \phi[x \mapsto y]$, where $y \notin \text{FV}(\phi) \cup \{x\}$ and $\phi[x \mapsto y]$ is safe. Then

$$\begin{aligned} \mathcal{M} \models_\alpha \exists y\, \phi[x \mapsto y] &\iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle y \mapsto a\rangle} \phi[x \mapsto y] \\ &\iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{(x \mapsto y)^{\mathcal{M}}(\alpha\langle y \mapsto a\rangle)} \phi \quad \text{by Lemma 4.41(b)} \\ &\iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle x \mapsto a\rangle} \phi \iff \mathcal{M} \models_\alpha \exists x\, \phi, \end{aligned}$$

since the assignment $(x \mapsto y)^{\mathcal{M}}(\alpha\langle y \mapsto a\rangle) : X \cup \{x\} \to M$ maps $x \mapsto y^{\mathcal{M}}(\alpha\langle y \mapsto a\rangle) = a$ and all other $z \in \text{FV}(\phi) \setminus \{x\}$ to $z^{\mathcal{M}}(\alpha\langle y \mapsto a\rangle) = \alpha(z)$ since $y \notin \text{FV}(\phi)$, hence agrees with $\alpha\langle x \mapsto a\rangle$ on $\text{FV}(\phi)$, and so the interpretations of ϕ under them are the same by Proposition 2.10.

Now we show by induction on the definition of $\approx_\alpha$ that if $\phi \approx_\alpha \psi$, then $\mathcal{M} \models_\alpha \phi \leftrightarrow \psi$:

- The base case is for $\sim_\alpha$, shown above.
- If $\phi \wedge \theta \approx_\alpha \psi \wedge \theta$ because $\phi \approx_\alpha \psi$, then

$$\begin{aligned} \mathcal{M} \models_\alpha \phi \wedge \theta &\iff \mathcal{M} \models_\alpha \phi \text{ and } \mathcal{M} \models_\alpha \theta \\ &\iff \mathcal{M} \models_\alpha \psi \text{ and } \mathcal{M} \models_\alpha \theta \quad \text{by IH} \\ &\iff \mathcal{M} \models_\alpha \psi \wedge \theta. \end{aligned}$$

- All the other cases (including $\exists$) are similar.

The claim for $\equiv_\alpha$ follows easily, since semantic equivalence is clearly transitive. □

Proof of Proposition 4.40. We assume that there is a deduction $\mathcal{D}$ of $\mathcal{T} \vdash_X \phi$, and we must show that for every $\mathcal{A}$-structure $\mathcal{M}$ and variable assignment $\alpha : X \to M$ such that $\mathcal{M} \models_\alpha \mathcal{T}$, we have $\mathcal{M} \models_\alpha \phi$. We use induction on $\mathcal{D}$.

- If $\mathcal{D}$ ends with (A), then $\phi \in \mathcal{T}$, so since $\mathcal{M} \models_\alpha \mathcal{T}$, we have $\mathcal{M} \models_\alpha \phi$.
- If $\mathcal{D}$ ends with a first-order instance of a propositional inference rule, then the same reasoning as in the proof of soundness for propositional logic (Proposition 3.22 in notes) applies.

  For example, if the deduction ends with

$$(\neg\text{E})\ \frac{\mathcal{T} \vdash_X \phi \qquad \mathcal{T} \vdash_X \neg\phi}{\mathcal{T} \vdash_X \bot},$$

  then for every $\mathcal{M}$ and $\alpha : X \to M$ such that $\mathcal{M} \models_\alpha \mathcal{T}$, by the IH, we have $\mathcal{M} \models_\alpha \phi$ and $\mathcal{M} \models_\alpha \neg\phi$, which is impossible; thus for every such $\mathcal{M}, \alpha$, we vacuously have $\mathcal{M} \models_\alpha \bot$.
- If $\mathcal{D}$ ends with

$$(=\text{I})\ \frac{}{\mathcal{T} \vdash_X t = t},$$

  we have $\mathcal{M} \models_\alpha t = t \iff t^{\mathcal{M}}(\alpha) = t^{\mathcal{M}}(\alpha)$ which is clearly true.
- If $\mathcal{D}$ ends with

$$(=\text{E})\ \frac{\mathcal{T} \vdash_X s = t \qquad \mathcal{T} \vdash_X \phi[x \mapsto s]}{\mathcal{T} \vdash_X \phi[x \mapsto t]}$$

  where $s, t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ and $\phi \in \mathcal{L}^{X\cup\{x\}}_{\text{form}}(\mathcal{A})$ with $\phi[x \mapsto s], \phi[x \mapsto t]$ safe, by the IH, we know

$$\begin{aligned}
\mathcal{M} \models_\alpha s = t &\iff s^{\mathcal{M}}(\alpha) = t^{\mathcal{M}}(\alpha), \\
\mathcal{M} \models_\alpha \phi[x \mapsto s] &\iff \mathcal{M} \models_{\alpha\langle x \mapsto s^{\mathcal{M}}(\alpha)\rangle} \phi \quad \text{by Lemma 4.41(b)} \\
&\iff \mathcal{M} \models_{\alpha\langle x \mapsto t^{\mathcal{M}}(\alpha)\rangle} \phi \quad \text{by above} \\
&\iff \mathcal{M} \models_\alpha \phi[x \mapsto t] \quad \text{by Lemma 4.41(b) again}
\end{aligned}$$

  (where we are using, in the second line for instance, that $(x \mapsto s)^{\mathcal{M}}(\alpha) = \alpha\langle x \mapsto s^{\mathcal{M}}(\alpha)\rangle$, since both map x to $s^{\mathcal{M}}(\alpha)$ and every $y \in X \setminus \{x\}$ to $\alpha(m)$).
- If $\mathcal{D}$ ends with

$$(\exists\text{I})\ \frac{\mathcal{T} \vdash_X \phi[x \mapsto t]}{\mathcal{T} \vdash_X \exists x\, \phi}$$

  with $\phi[x \mapsto t]$ safe, then by the IH, we have $\mathcal{M} \models_\alpha \phi[x \mapsto t]$, which by Lemma 4.41(b) means $\mathcal{M} \models_{\alpha\langle x \mapsto t^{\mathcal{M}}(\alpha)\rangle} \phi$; thus there is $a \in M$ such that $\mathcal{M} \models_{\alpha\langle x \mapsto a\rangle} \phi$, i.e., $\mathcal{M} \models_\alpha \exists x\, \phi$.
- Finally, suppose $\mathcal{D}$ ends with

$$(\exists\text{E})\ \frac{\mathcal{T} \vdash_X \exists x\, \phi \qquad \mathcal{T} \cup \{\phi\} \vdash_{X\cup\{x\}} \psi}{\mathcal{T} \vdash_X \psi}$$

  with $x \notin X$. By the first IH, we know

$$\mathcal{M} \models_\alpha \exists x\, \phi \iff \exists a \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle x \mapsto a\rangle} \phi.$$

  Since also $\mathcal{M} \models_\alpha \mathcal{T}$ by assumption, and so $\mathcal{M} \models_{\alpha\langle x \mapsto a\rangle} \mathcal{T}$ (by Proposition 2.10) since $\mathcal{T}$ only has free variables from $X \not\ni x$, by the second IH, we get

$$\mathcal{M} \models_{\alpha\langle x \mapsto a\rangle} \psi.$$

  Since ψ also only has free variables from $X \not\ni x$, by Proposition 2.10 again, this means

$$\mathcal{M} \models_\alpha \psi,$$

  as desired. □

5. Completeness

Let $\mathcal{A}$ be a first-order signature, $\mathcal{T}$ be an open $\mathcal{A}$-theory, and ϕ be an $\mathcal{A}$-formula, both with free variables from X. By soundness (Proposition 4.40), if $\mathcal{T} \vdash_X \phi$, then $\mathcal{T} \models_X \phi$, i.e., "provable statements are true in all models", where here a "model" of the open theory $\mathcal{T}$ consists of an $\mathcal{A}$-structure $\mathcal{M}$ *together with* a variable assignment $\alpha : X \to M$ satisfying all the formulas in $\mathcal{T}$.

Theorem 5.1 (completeness)**.** If $\mathcal{T} \models_X \phi$, then $\mathcal{T} \vdash_X \phi$.

Our proof strategy will be an extension of what we did in propositional logic. Suppose $\mathcal{T} \nvdash_X \phi$. We will show that $\mathcal{T} \not\models_X \phi$, i.e., we will construct an $\mathcal{A}$-structure $\mathcal{M}$ together with a variable assignment $\alpha : X \to M$ such that $\mathcal{M} \models_\alpha \mathcal{T}$ but $\mathcal{M} \not\models_\alpha \phi$. We would like to define $\mathcal{M}$ to consist of "exactly what the theory $\mathcal{T}$ demands" (similar to what we did in propositional logic); we will see that there are three conditions on $\mathcal{T}$ that ensure we are able to do so (namely, consistency, completeness, and an additional "witness property"). To finish the proof, we will show that any $\mathcal{T}$ may be extended to some $\mathcal{T}' \supseteq \mathcal{T}$ obeying these conditions.

To define $\mathcal{M}$, we must first specify its underlying set M. In order to get an $\mathcal{A}$-structure $\mathcal{M}$ with a variable assignment $\alpha : M \to X$, each term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ will need to have an interpretation $t^{\mathcal{M}}(\alpha) \in M$. Thus, as a first approximation, we might take the underlying set to be simply the set of terms $\mathcal{L}^X_{\text{term}}(\mathcal{A})$, where we think of a term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ as its own interpretation. However, the theory $\mathcal{T}$ requires some terms to have the same interpretation: for example,

$$\mathcal{T}_{\mathsf{abgrp}} \vdash_{\{x,y\}} x + y = y + x$$

(by a simple application of ($\forall$E)), and so any model $\mathcal{M} \models \mathcal{T}_{\mathsf{abgrp}}$ will have to interpret $x + y, y + x$ as the same element, by soundness. We therefore take

$$M := \mathcal{L}^X_{\text{term}}(\mathcal{A})/{\equiv_{\mathcal{T}}},$$

where $\equiv_{\mathcal{T}}$ is the **$\mathcal{T}$-provable equality** relation between terms defined by

$$s \equiv_{\mathcal{T}} t \ :\iff \ \mathcal{T} \vdash_X s = t.$$

In other words, M consists of elements which have to exist in any $\mathcal{A}$-structure with a variable assignment $\alpha : X \to M$, which are equal precisely when $\mathcal{T}$ says they have to be.

Lemma 5.2. $\equiv_{\mathcal{T}}$ is an equivalence relation on $\mathcal{L}^X_{\text{term}}(\mathcal{A})$.

Proof. By the (=I), (Sym), and (Trans) rules. □

Lemma 5.3. Let $s_1, \dots, s_n, t_1, \dots, t_n \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ be terms. If $s_1 \equiv_{\mathcal{T}} t_1$, …, $s_n \equiv_{\mathcal{T}} t_n$, then:

(a) For each function symbol $f \in \mathcal{A}^n_{\text{fun}}$, we have $f(s_1, \dots, s_n) \equiv_{\mathcal{T}} f(t_1, \dots, t_n)$; thus

$$f^{\mathcal{M}} : M^n = (\mathcal{L}^X_{\text{term}}(\mathcal{A})/{\equiv_{\mathcal{T}}})^n \longrightarrow \mathcal{L}^X_{\text{term}}(\mathcal{A})/{\equiv_{\mathcal{T}}} = M$$
$$([t_1], \dots, [t_n]) \longmapsto [f(t_1, \dots, t_n)].$$

is a well-defined function.

(b) For each $R \in \mathcal{A}^n_{\text{rel}}$, we have $\mathcal{T} \vdash_X R(s_1, \dots, s_n) \iff \mathcal{T} \vdash_X R(t_1, \dots, t_n)$; thus

$$R^{\mathcal{M}} : M^n = (\mathcal{L}^X_{\text{term}}(\mathcal{A})/{\equiv_{\mathcal{T}}})^n \longrightarrow \{0, 1\}$$
$$([t_1], \dots, [t_n]) \longmapsto \begin{cases} 1 & \text{if } \mathcal{T} \vdash_X R(t_1, \dots, t_n), \\ 0 & \text{otherwise} \end{cases}$$

is well-defined.

Proof. (a) is by the (Cong) rule. The proof of (b) is similar to the proof of the (Cong) rule in Example 4.15: if $\mathcal{T} \vdash_X R(s_1, \dots, s_n)$, we get $\mathcal{T} \vdash_X R(t_1, \dots, t_n)$ by repeatedly applying (=E) with the deductions of $\mathcal{T} \vdash_X s_1 = t_1$, …, coming from $s_1 \equiv_{\mathcal{T}} t_1$, …. □

We have now defined a structure $\mathcal{M}$ from an arbitrary open theory $\mathcal{T} \subseteq \mathcal{L}^X_{\text{form}}(\mathcal{A})$, consisting of $\mathcal{T}$-provable equivalence classes of terms. Under the following variable assignment

$$\begin{aligned} \alpha : X &\longrightarrow M \\ x &\longmapsto [x], \end{aligned}$$

we claim that indeed, the interpretation of each term in $\mathcal{M}$ is "itself" (or rather, its $\mathcal{T}$-equivalence class):

Lemma 5.4. For any $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$, we have

$$t^{\mathcal{M}}(\alpha) = [t].$$

Proof. By induction on t.

- For a variable $t = x \in X$, we have $x^{\mathcal{M}}(\alpha) = \alpha(x) = [x]$.
- For $t = f(t_1, \ldots, t_n)$ where $f \in \mathcal{A}^n_{\text{fun}}$, we have

$$\begin{aligned} f(t_1, \ldots, t_n)^{\mathcal{M}}(\alpha) &= f^{\mathcal{M}}(t_1^{\mathcal{M}}(\alpha), \ldots, t_n^{\mathcal{M}}(\alpha)) \\ &= f^{\mathcal{M}}([t_1], \ldots, [t_n]) \quad \text{by IH} \\ &= [f(t_1, \ldots, t_n)] \qquad \text{by definition of } f^{\mathcal{M}}. \end{aligned}$$

□

The counterpart of Lemma 5.4 for formulas is the next Lemma 5.8, which says that the structure $\mathcal{M}$ satisfies "exactly what the theory $\mathcal{T}$ demands". The proof of this is analogous to the proof for propositional logic (Lemma 4.2, and the discussion preceding it, in the notes). As in that proof, in order for the induction to work, we need $\mathcal{T}$ to be

- **consistent**: $\mathcal{T} \nvdash_X \bot$ (equivalently by ($\bot$E), $\mathcal{T} \nvdash_X \phi$ for some $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$);
- **complete**: for all $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, either $\mathcal{T} \vdash_X \phi$ or $\mathcal{T} \vdash_X \neg\phi$;

recall that these properties are used in the $\bot$ ("0-ary $\vee$") and $\vee$ cases, respectively. Recalling the analogy between $\exists$ and $\vee$ (Remark 4.18), it is no surprise that in first-order logic, we also need $\mathcal{T}$ to obey a third condition, namely

- the **witness property**: if $\mathcal{T} \vdash_X \exists x\, \phi$, then there is some term $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ and $\phi' \equiv_\alpha \phi$ such that $\phi'[x \mapsto t]$ is safe and $\mathcal{T} \vdash_X \phi'[x \mapsto t]$.

Note that conversely, if $\mathcal{T} \vdash_X \phi'[x \mapsto t]$, then $\mathcal{T} \vdash_X \exists x\, \phi' \equiv_\alpha \exists x\, \phi$ by ($\exists$I). Thus, the witness property can be seen as saying that $\mathcal{T}$ obeys a "converse" of ($\exists$I).

Example 5.5. We have $\{\exists x\, \top\} \vdash_\varnothing \exists x\, \top$ by (A), but there is no term with free variables from $\varnothing$ (assuming the signature $\mathcal{A} = \varnothing$); thus $\mathcal{T} := \{\exists x\, \top\}$ does not have the witness property over $\varnothing$.

Example 5.6. We have $\mathcal{T}_{\mathsf{ordfield}} \vdash \exists x\, (x + x = 1)$, but there is no closed $\mathcal{L}_{\mathsf{ordfield}}$-term denoting $1/2$, so $\mathcal{T}_{\mathsf{ordfield}}$ does not have the witness property.

Exercise 5.7. Verify this.
[Hint: to prove $\exists x\, (x + x = 1)$, the key point is that $\mathcal{T}_{\mathsf{ordfield}} \vdash \neg(0 = 1 + 1)$; see Example 4.11, Extra Practice Problem 2.6(c). To prove that there is no closed witness term for $\exists x\, (x + x = 1)$, consider its interpretation in $\mathbb{R}$, say.]

Lemma 5.8. Let $\mathcal{T}$ be an open theory with free variables from X, and let $\mathcal{M}$ and $\alpha : X \to M$ be the structure and variable assignment defined above. Then $\mathcal{T}$ is consistent and complete and has the witness property iff for all $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, we have

$$(*) \qquad \mathcal{M} \models_\alpha \phi \iff \mathcal{T} \vdash_X \phi.$$

Proof. ($\Longleftarrow$) Since $\mathcal{M} \not\models_\alpha \bot$, $\mathcal{T} \not\vdash_X \bot$. For any $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$, either $\mathcal{M} \models_\alpha \phi$ or $\mathcal{M} \models_\alpha \neg\phi$; thus either $\mathcal{T} \vdash_X \phi$ or $\mathcal{T} \vdash_X \neg\phi$. If $\mathcal{T} \vdash_X \exists x\, \phi$, then by soundness, we know that

$$\begin{aligned}\mathcal{M} \models_\alpha \exists x\, \phi &\iff \exists [t] \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle x \mapsto [t]\rangle} \phi \\ &\iff \exists [t] \in M \text{ s.t. } \mathcal{M} \models_{(x\mapsto t)^{\mathcal{M}}(\alpha)} \phi,\end{aligned}$$

since both variable assignments above map $x \mapsto t^{\mathcal{M}}(\alpha) = [t]$ by Lemma 5.4 and $y \in X \setminus \{x\}$ to $\alpha(y)$; now by Lemmas 4.41 and 4.42 (and 3.25), we have

$$\begin{aligned}&\iff \exists [t] \in M,\ \phi' \equiv_\alpha \phi \text{ s.t. } \phi'[x \mapsto t] \text{ is safe and } \mathcal{M} \models_{(x\mapsto t)^{\mathcal{M}}(\alpha)} \phi' \\ &\iff \exists [t] \in M,\ \phi' \equiv_\alpha \phi \text{ s.t. } \phi'[x \mapsto t] \text{ is safe and } \mathcal{M} \models_\alpha \phi'[x \mapsto t] \\ &\iff \exists [t] \in M,\ \phi' \equiv_\alpha \phi \text{ s.t. } \phi'[x \mapsto t] \text{ is safe and } \mathcal{T} \vdash_X \phi'[x \mapsto t]\end{aligned}$$

by $(*)$, which proves that $\mathcal{T}$ has the witness property.

($\Longrightarrow$) This is mostly by induction on ϕ, except that in the $\exists x\, \phi$ case, we will need to use the IH not just for ϕ but for an arbitrary substitution $\phi[x \mapsto t]$. Thus, we really need to perform induction on the height $\mathrm{HT}(\phi)$, as defined in Exercise 3.29 and in such a way (see Exercise 3.29) that $\phi, \phi[x \mapsto t]$ have the same height for any *term* (not just variable) t.

- For atomic $\phi = R(t_1, \ldots, t_n)$ where $R \in \mathcal{A}^n_{\text{rel}}$, (similarly to the inductive case in Lemma 5.4)

$$\begin{aligned}\mathcal{M} \models_\alpha R(t_1, \ldots, t_n) &\iff R^{\mathcal{M}}(t_1^{\mathcal{M}}(\alpha), \ldots, t_n^{\mathcal{M}}(\alpha)) = 1 \\ &\iff R^{\mathcal{M}}([t_1], \ldots, [t_n]) = 1 \quad \text{by Lemma 5.4} \\ &\iff \mathcal{T} \vdash_X R(t_1, \ldots, t_n) \quad \text{by definition of } R^{\mathcal{M}}.\end{aligned}$$

- For atomic $\phi = (s = t)$,

$$\begin{aligned}\mathcal{M} \models_\alpha s = t &\iff s^{\mathcal{M}}(\alpha) = t^{\mathcal{M}}(\alpha) \\ &\iff [s] = [t] \quad \text{by Lemma 5.4} \\ &\iff \mathcal{T} \vdash_X s = t \quad \text{by definition of } \equiv_{\mathcal{T}}.\end{aligned}$$

- The connective cases are the same as in the proof for propositional logic (Lemma 4.2, and the discussion preceding it, in the notes). For example, if $(*)$ holds for ϕ, then to prove that it also holds for $\neg\phi$:

$$\begin{aligned}\mathcal{M} \models_\alpha \neg\phi &\iff \mathcal{M} \not\models_\alpha \phi \\ &\iff \mathcal{T} \not\vdash_X \phi \quad \text{by IH} \\ &\iff \mathcal{T} \vdash_X \neg\phi \quad \text{by consistency and completeness of } \mathcal{T}.\end{aligned}$$

- Finally, suppose $(*)$ holds for all formulas of height $< \mathrm{HT}(\exists x\, \phi)$, where $\phi \in \mathcal{L}^{X\cup\{x\}}_{\text{form}}(\mathcal{A})$; in particular, it holds for all safe $\phi'[x \mapsto t]$ where $t \in \mathcal{L}^X_{\text{term}}(\mathcal{A})$ and $\phi' \equiv_\alpha \phi$ (by Exercise 3.29). Then as in the proof of ($\Longrightarrow$) above, we have

$$\begin{aligned}&\mathcal{M} \models_\alpha \exists x\, \phi \\ &\iff \exists [t] \in M \text{ s.t. } \mathcal{M} \models_{\alpha\langle x \mapsto [t]\rangle} \phi \\ &\iff \exists [t] \in M \text{ s.t. } \mathcal{M} \models_{(x\mapsto t)^{\mathcal{M}}(\alpha)} \phi \\ &\iff \exists [t] \in M,\ \phi' \equiv_\alpha \phi \text{ s.t. } \phi'[x \mapsto t] \text{ is safe and } \mathcal{M} \models_{(x\mapsto t)^{\mathcal{M}}(\alpha)} \phi' \quad \text{by Lemma 4.42} \\ &\iff \exists [t] \in M,\ \phi' \equiv_\alpha \phi \text{ s.t. } \phi'[x \mapsto t] \text{ is safe and } \mathcal{M} \models_\alpha \phi'[x \mapsto t] \quad \text{by Lemma 4.41} \\ &\iff \exists [t] \in M,\ \phi' \equiv_\alpha \phi \text{ s.t. } \phi'[x \mapsto t] \text{ is safe and } \mathcal{T} \vdash_X \phi'[x \mapsto t] \quad \text{by IH} \\ &\iff \mathcal{T} \vdash_X \exists x\, \phi\end{aligned}$$

by ($\exists$I) and the witness property for $\mathcal{T}$. □

Now if $\mathcal{T}$ is a complete theory with the witness property such that $\mathcal{T} \nvdash_X \phi$ (hence $\mathcal{T}$ is consistent), then by Lemma 5.8, we have $\mathcal{M} \models_\alpha \mathcal{T}$ but $\mathcal{M} \not\models_\alpha \phi$, whence $\mathcal{T} \not\models_X \phi$. So to finish the proof of the completeness theorem, we need to modify an arbitrary theory $\mathcal{T}$ to give it these properties.

- As in propositional logic, completeness will be achieved by repeatedly adding axioms to $\mathcal{T}$ until it becomes complete.
- In order to achieve the witness property, whenever $\mathcal{T} \vdash_X \exists x\, \phi$, we will add a new *variable* to X, which will serve as a witness for $\exists x\, \phi$. There will now be new formulas involving the new variable, so we will need to repeat this step (as well as the previous step) in order to fix the conditions for the new formulas.

This procedure is formalized as follows.

Lemma 5.9. Let $\mathcal{T}$ be an open theory and ϕ be a formula, both with free variables from X, such that $\mathcal{T} \nvdash_X \phi$. Then there is a theory $\mathcal{T}' \supseteq \mathcal{T}$ with free variables from some $X' \supseteq X$, which is complete and has the witness property (for formulas over X'), such that $\mathcal{T}' \nvdash_{X'} \phi$.

Proof of completeness theorem given Lemma 5.9. Suppose $\mathcal{T} \nvdash_X \phi$. Then by Lemma 5.9, there is $\mathcal{T}' \supseteq \mathcal{T}$ with free variables from $X' \supseteq X$, which is complete and has the witness property, such that $\mathcal{T}' \nvdash_{X'} \phi$, whence $\mathcal{T}'$ is also consistent. By Lemma 5.8, we get an $\mathcal{A}$-structure $\mathcal{M}$ together with a variable assignment $\alpha : X' \to M$ such that

$$\mathcal{M} \models_\alpha \psi \iff \mathcal{T}' \vdash_{X'} \psi$$

for all $\psi \in \mathcal{L}^{X'}_{\text{form}}(\mathcal{A})$. In particular, for all $\psi \in \mathcal{T} \subseteq \mathcal{T}'$, we have $\mathcal{T}' \vdash_{X'} \psi$ (by (A)) so $\mathcal{M} \models_\alpha \psi$ and so $\mathcal{M} \models_{\alpha|X} \psi$ since $\text{FV}(\psi) \subseteq X$ (using Proposition 2.10), i.e., $\mathcal{M} \models_{\alpha|X} \mathcal{T}$; and since $\mathcal{T}' \nvdash_{X'} \phi$, we have $\mathcal{M} \not\models_\alpha \phi$, so again since $\text{FV}(\phi) \subseteq X$, $\mathcal{M} \not\models_{\alpha|X} \phi$. So $\mathcal{M}, \alpha|X$ witnesses that $\mathcal{T} \not\models_X \phi$. □

To prove Lemma 5.9, we need to know: (1) we can add a single axiom to $\mathcal{T}$; (2) we can add a new *variable* to serve as a witness for an existential; and (3) we can repeat both of these steps.

Lemma 5.10. Let $\mathcal{T} \nvdash_X \phi$, and let $\psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ be another formula. Then either $\mathcal{T} \cup \{\psi\} \nvdash_X \phi$ or $\mathcal{T} \cup \{\neg\psi\} \nvdash_X \phi$.

Proof. As in propositional logic (Lemma 4.4 in the notes). □

Lemma 5.11. Let $\mathcal{T} \nvdash_X \phi$, and let $\psi \in \mathcal{L}^{X \cup \{x\}}_{\text{form}}(\mathcal{A})$ such that $\mathcal{T} \vdash_X \exists x\, \psi$. Then there is a variable y such that $\psi[x \mapsto y]$ is safe and $\mathcal{T} \cup \{\psi[x \mapsto y]\} \nvdash_{X \cup \{y\}} \phi$.

Proof. Let $y \notin X \cup \{x\} \cup \text{BV}(\psi)$, so that $\exists x\, \psi \sim_\alpha \exists y\, \psi[x \mapsto y]$ (using Exercise 3.22). If we had $\mathcal{T} \cup \{\psi[x \mapsto y]\} \vdash_{X \cup \{y\}} \phi$, then by ($\exists$E) applied to $\mathcal{T} \vdash_X \exists x\, \psi$, we would have $\mathcal{T} \vdash_X \phi$. □

Lemma 5.12. Let $\mathcal{T}_0 \subseteq \mathcal{T}_1 \subseteq \cdots$ be an increasing sequence of open theories, with free variables from $X_0 \subseteq X_1 \subseteq \cdots$ respectively, such that $\mathcal{T}_n \nvdash_{X_n} \phi$ for each n. Then $\bigcup_n \mathcal{T}_n \nvdash_{\bigcup_n X_n} \phi$.

Proof. Suppose $\bigcup_n \mathcal{T}_n \vdash_{\bigcup_n X_n} \phi$. By syntactic compactness (Proposition 4.36), there are finite $\mathcal{T}' \subseteq \bigcup_n \mathcal{T}_n$ and $X' \subseteq \bigcup_n X_n$ such that $\mathcal{T}' \vdash_{X'} \phi$. Since $\mathcal{T}', X'$ are finite, there is some n such that $\mathcal{T}' \subseteq \mathcal{T}_n$ and $X' \subseteq X_n$, whence by variable weakening (Corollary 4.35), $\mathcal{T}' \vdash_{X_n} \phi$, and then by weakening, $\mathcal{T}_n \vdash_{X_n} \phi$. □

We can now repeat step (1) to achieve completeness:

Lemma 5.13. Let $\mathcal{T} \nvdash_X \phi$. Then there is a complete theory $\mathcal{T}' \supseteq \mathcal{T}$, still with free variables from X, such that $\mathcal{T}' \nvdash_X \phi$.

Proof. The proof is the same as in propositional logic (Lemma 4.3 in notes): enumerate $\mathcal{L}^X_{\text{form}}(\mathcal{A})$, and for each formula, add either it or its negation to $\mathcal{T}$ using Lemma 5.10, then take the union of these theories and use Lemma 5.12. (If $\mathcal{L}^X_{\text{form}}(\mathcal{A})$ is uncountable, use either transfinite induction or Zorn's lemma.) □

Next, we use step (2) to achieve the witness property *for all formulas with the original free variables*, after which we need to repeat both steps (1) and (2) to handle the newly added variables:

Lemma 5.14. Let $\mathcal{T} \nvdash_X \phi$. Then there is $\mathcal{T}' \supseteq \mathcal{T}$ with free variables from some $X' \supseteq X$ such that $\mathcal{T}' \nvdash_{X'} \phi$, and $\mathcal{T}'$ has the witness property for all existential formulas $\exists x\, \psi$ which are proved by $\mathcal{T}$ over X (rather than by $\mathcal{T}'$ over X').

Proof. The proof is similar to the previous proof, using Lemma 5.11 to extend $\mathcal{T}, X$ for each possible $\exists x\, \psi$ with free variables from X. If $\mathcal{L}_{\text{form}}^{X\cup\{x\}}(\mathcal{A})$ is countable, enumerate $\mathcal{L}_{\text{form}}^{X\cup\{x\}}(\mathcal{A}) = \{\psi_0, \psi_1, \dots\}$, and inductively define an increasing sequence of theories $\mathcal{T}_0 \subseteq \mathcal{T}_1 \subseteq \cdots$ with free variables from $X_0 \subseteq X_1 \subseteq \cdots \subseteq \mathcal{V}$ respectively, so that $\mathcal{T}_n \nvdash_{X_n} \phi$ for each n, as follows:

- Let $\mathcal{T}_0 := \mathcal{T}$ and $X_0 := X$.
- Given $\mathcal{T}_n$ and X_n, if $\mathcal{T}_n \nvdash_{X_n} \exists x\, \psi_n$, let $\mathcal{T}_{n+1} := \mathcal{T}_n$ and $X_{n+1} := X_n$. Otherwise, by Lemma 5.11, there is a variable y_n such that $\psi_n[x \mapsto y_n]$ is safe and
$$\mathcal{T}_{n+1} := \mathcal{T}_n \cup \{\psi_n[x \mapsto y_n]\} \nvdash_{X_{n+1} := X_n \cup \{y_n\}} \phi.$$

Now let
$$\mathcal{T}' := \bigcup_n \mathcal{T}_n, \qquad X' := \bigcup_n X_n.$$
Then

- $\mathcal{T}'$ has the witness property for all $\exists x\, \psi$ proved by $\mathcal{T}$ over X, since ψ must be ψ_n for some n, whence $\mathcal{T} \vdash_X \exists x\, \psi_n$ implies $\mathcal{T}_n \vdash_{X_n} \exists x\, \psi_n$ by (variable) weakening, so by definition of $\mathcal{T}_{n+1}, X_{n+1}$, we have $y_n \in X_{n+1} \subseteq X'$ and (safe) $\psi_n[x \mapsto y_n] \in \mathcal{T}_{n+1} \subseteq \mathcal{T}'$ whence $\mathcal{T}' \vdash_{X'} \psi_n[x \mapsto y_n]$ by (A).
- $\mathcal{T}' \nvdash_{X'} \phi$ by Lemma 5.12, since $\mathcal{T}_0 = \mathcal{T} \nvdash_{X_0 = X} \phi$, so by induction, $\mathcal{T}_n \nvdash_{X_n} \phi$ for each n.

If $\mathcal{L}_{\text{form}}^{X\cup\{x\}}(\mathcal{A})$ is uncountable, use either transfinite induction or Zorn's lemma. □

Proof of Lemma 5.9. Define an increasing sequence of theories $\mathcal{T}_0 \subseteq \mathcal{T}_1 \subseteq \cdots$, with free variables from $X_0 \subseteq X_1 \subseteq \cdots \subseteq \mathcal{V}$, so that $\mathcal{T}_n \nvdash_{X_n} \phi$ for each n, by induction as follows:

- Let $\mathcal{T}_0 := \mathcal{T}$ and $X_0 := X$.
- Given $\mathcal{T}_n$ and X_n, by Lemma 5.13, there is complete $\mathcal{T}_n' \supseteq \mathcal{T}_n$ with free variables from X_n such that $\mathcal{T}_n' \nvdash_{X_n} \phi$, and then by Lemma 5.14, there is $\mathcal{T}_{n+1} \supseteq \mathcal{T}_n'$ with free variables from $X_{n+1} \supseteq X_n$ which has the witness property for all existentials proved by $\mathcal{T}_n'$ over X_n and still satisfies $\mathcal{T}_{n+1} \nvdash_{X_{n+1}} \phi$.

Let
$$\mathcal{T}' := \bigcup_n \mathcal{T}_n, \qquad X' := \bigcup_n X_n.$$
Then

- $\mathcal{T}'$ is complete (over X'), since for any $\psi \in \mathcal{L}_{\text{form}}^{X'}(\mathcal{A})$, we have $\text{FV}(\psi) \subseteq X_n$ for some n, whence by completeness of $\mathcal{T}_n'$, either $\mathcal{T}_n' \vdash_{X_n} \psi$ or $\mathcal{T}_n' \vdash_{X_n} \neg\psi$, and so by (variable) weakening, either $\mathcal{T}' \vdash_{X'} \psi$ or $\mathcal{T}' \vdash_{X'} \neg\psi$.
- $\mathcal{T}'$ has the witness property (over X'), since for any $\exists x\, \psi \in \mathcal{L}_{\text{form}}^{X'}(\mathcal{A})$ such that $\mathcal{T}' \vdash_{X'} \exists x\, \psi$, by syntactic compactness, we have $\mathcal{T}_n \vdash_{X_n} \exists x\, \psi$ for some n, whence $\mathcal{T}_n' \vdash_{X_n} \exists x\, \psi$ by weakening, so since $\mathcal{T}_{n+1}$ has the witness property for existentials proved by $\mathcal{T}_n'$ over X_n, there is $t \in \mathcal{L}_{\text{term}}^{X_{n+1}}(\mathcal{A})$ and $\psi' \equiv_\alpha \psi$ with $\psi'[x \mapsto t]$ safe and $\mathcal{T}_{n+1} \vdash_{X_{n+1}} \psi'[x \mapsto t]$, so by (variable) weakening, $\mathcal{T}' \vdash_{X'} \psi'[x \mapsto t]$.
- $\mathcal{T}' \nvdash_{X'} \phi$ by Lemma 5.12.

(Note that there is no need to go into the transfinite here, since we're not enumerating formulas.) □

This concludes the proof of the completeness theorem for first-order logic.

5.1. **Consequences of completeness.** Soundness and completeness together say

Corollary 5.15. For any open theory $\mathcal{T}$ and formula ϕ with free variables from X, we have

$$\mathcal{T} \vdash_X \phi \iff \mathcal{T} \models_X \phi.$$

In particular (taking $\phi = \bot$), $\mathcal{T}$ is consistent iff it is satisfiable, i.e., has a model.

Here, as at the start of Section 5, a "model" of an open theory $\mathcal{T} \subseteq \mathcal{L}^X_{\text{form}}(\mathcal{A})$ should be read as meaning an $\mathcal{A}$-structure $\mathcal{M}$ *together with* a variable assignment $\alpha : X \to M$ with $\mathcal{M} \models_\alpha \mathcal{T}$. Let

$$\text{Mod}_X(\mathcal{T}) := \{(\mathcal{M}, \alpha) \mid \mathcal{M} \text{ is an } \mathcal{A}\text{-structure}, \alpha : X \to M, \mathcal{M} \models_\alpha \mathcal{T}\}.$$

When $X = \varnothing$, this agrees with $\text{Mod}(\mathcal{T})$ for a closed theory $\mathcal{T}$ as defined in Section 2.3. The above corollary then says that $\mathcal{T}$ is consistent (over X) iff $\text{Mod}_X(\mathcal{T}) \neq \varnothing$.

Unlike in propositional logic, it is far from true that $\mathcal{T}$ is complete iff it has at most one model. Indeed, we already know from Section 2.5 that pretty much every nontrivial theory has many models: if $\mathcal{M} \models \mathcal{T}$, then any isomorphic copy $\mathcal{N} \cong \mathcal{M}$ will also be a model of $\mathcal{T}$; more generally, any $\mathcal{N}$ admitting an elementary embedding to/from $\mathcal{M}$ will also be a model of $\mathcal{T}$. In other words, if we define for each structure $\mathcal{M}$ its **complete theory**

$$\text{Th}(\mathcal{M}) := \{\phi \in \mathcal{L}^\varnothing_{\text{form}}(\mathcal{A}) \mid \mathcal{M} \models \phi\}$$

consisting of all first-order expressible properties of $\mathcal{M}$, similarly to in propositional logic (see Proposition 4.12 in notes), $\text{Th}(\mathcal{M})$ will usually fail badly to uniquely determine $\mathcal{M}$; there will usually be many other $\mathcal{N} \models \text{Th}(\mathcal{M})$, which look the same as $\mathcal{M}$ as far as first-order sentences can tell. Note that $\text{Th}(\mathcal{M})$ is indeed a complete (as well as consistent) (closed) theory.

Call two $\mathcal{A}$-structures $\mathcal{M}, \mathcal{N}$ **elementarily equivalent** if $\text{Th}(\mathcal{M}) = \text{Th}(\mathcal{N})$, or equivalently (by considering negations of sentences), $\text{Th}(\mathcal{M}) \subseteq \text{Th}(\mathcal{N})$, i.e., $\mathcal{N} \models \text{Th}(\mathcal{M})$. Thus,

$$\text{isomorphic} \implies \text{elementarily equivalent};$$

more generally,

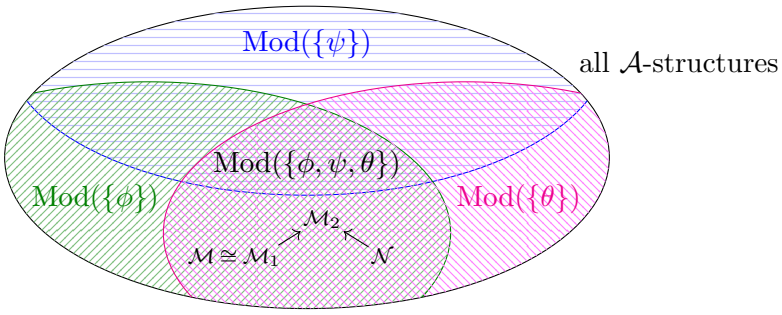
$$\exists \text{ elementary embedding } \mathcal{M} \to \mathcal{N} \implies \mathcal{M}, \mathcal{N} \text{ elementarily equivalent}.$$

Note that an elementary embedding is an *asymmetric* notion, whereas elementary equivalence is of course an equivalence relation (on the collection of all $\mathcal{A}$-structures). It follows that if we "symmetrize" elementary embeddings by allowing a finite zigzag of them, the result still implies elementary equivalence:

$$\exists \text{ elementary embeddings } \mathcal{M} \to \mathcal{M}_1 \leftarrow \mathcal{M}_2 \to \cdots \leftarrow \mathcal{N} \implies \mathcal{M}, \mathcal{N} \text{ elementarily equivalent}.$$

In fact, the converse of this last implication is true as well: if $\mathcal{M}, \mathcal{N}$ are elementarily equivalent, then there has to be a zigzag of elementary embeddings between them (of length ≤ 2). However, this is a nontrivial theorem whose proof requires the compactness theorem; see Theorem 5.25 below.

The following Venn diagram, analogous to the one we drew for propositional logic (Remark 4.13 in notes), illustrates the above notions:

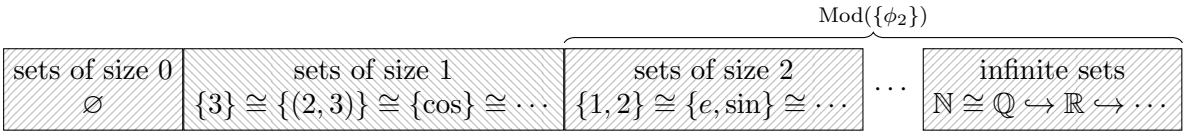


Each sentence ϕ picks out the subcollection $\mathrm{Mod}(\{\phi\})$ of all $\mathcal{A}$-structures which satisfy ϕ; for a theory $\mathcal{T}$, $\mathrm{Mod}(\mathcal{T})$ is then the intersection of these for all $\phi \in \mathcal{T}$. For a structure $\mathcal{M}$, its complete theory $\mathrm{Th}(\mathcal{M})$ consists of all those ϕ whose oval contains $\mathcal{M}$; in the picture above, $\mathrm{Th}(\mathcal{M})$ would contain $\phi, \neg\psi, \theta$ (as well as many other sentences). The elementary equivalence class $\mathrm{Mod}(\mathrm{Th}(\mathcal{M}))$ of $\mathcal{M}$ is then the intersection of all $\mathrm{Mod}(\{\phi\})$'s which contain $\mathcal{M}$. If two $\mathcal{M}, \mathcal{N}$ are linked by a finite chain of elementary embedding arrows (in either direction, including isomorphisms), then they must fall on the same side of every $\mathrm{Mod}(\{\phi\})$, hence belong to the same elementary equivalence class. Conversely, Theorem 5.25 below says that if two $\mathcal{M}, \mathcal{N}$ are *not* linked by a finite chain of arrows, then they must be separated by *some* $\mathrm{Mod}(\{\phi\})$.

Example 5.16. Consider $\mathcal{A} = \varnothing$, whose structures are just sets. By Exercise 2.64, all infinite sets are elementarily equivalent; indeed, its proof from HW7 explicitly shows how two infinite sets may be linked by a chain of two elementary embeddings (i.e., injections). On the other hand, for $n \in \mathbb{N}$,

$$\phi_n := \exists x_1 \cdots \exists x_n \left(\textstyle\bigwedge_{i<j} \neg(x_i = x_j)\right)$$

is a sentence axiomatizing the sets with $\geq n$ elements; thus $\phi_n \wedge \neg\phi_{n+1}$ is a sentence axiomatizing the sets with exactly n elements. So the elementary equivalence classes of sets look like:

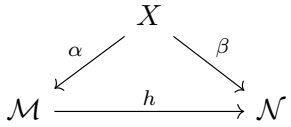


For each finite n, we have an elementary equivalence class of sets of size n, which are all isomorphic; except when $n = 0$, there are many (indeed, a proper class of) distinct such sets of size n. We also have a single elementary equivalence class consisting of all infinite sets, which are linked by elementary embeddings (i.e., injections), but are *not* all isomorphic, since their cardinalities vary.

The above notions admit obvious generalizations when free variables $X \neq \varnothing$ are allowed. For a structure $\mathcal{M}$ *together with* a variable assignment $\alpha : X \to \mathcal{M}$, define its **complete theory** to be

$$\mathrm{Th}_X(\mathcal{M}, \alpha) := \{\phi \in \mathcal{L}^X_{\mathrm{form}}(\mathcal{A}) \mid \mathcal{M} \models_\alpha \phi\}.$$

Call such $(\mathcal{M}, \alpha)$ and $(\mathcal{N}, \beta)$ **elementarily equivalent** if $\mathrm{Th}_X(\mathcal{M}, \alpha) = \mathrm{Th}_X(\mathcal{N}, \beta)$, or equivalently $\mathcal{N} \models_\beta \mathrm{Th}_X(\mathcal{M}, \alpha)$. The corresponding notion of **isomorphism** or **elementary embedding** $h : (\mathcal{M}, \alpha) \to (\mathcal{N}, \beta)$ between two structures-with-variable-assignments would be an isomorphism or elementary embedding $h : \mathcal{M} \to \mathcal{N}$ such that $h \circ \alpha = \beta$:



By definition of preservation of formulas (see Section 2.5), this ensures that

$$\mathcal{M} \models_\alpha \phi \iff \mathcal{N} \models_{h\circ\alpha} \phi \iff \mathcal{N} \models_\beta \phi,$$

so that again, the existence of an elementary embedding implies elementary equivalence.

5.2. **Compactness.** The following follows from syntactic compactness (Proposition 4.36) via soundness and completeness:

Corollary 5.17 (compactness)**.** If $\mathcal{T} \models_X \phi$, then there are finite $\mathcal{T}' \subseteq \mathcal{T}$ and $X' \subseteq X$ containing all free variables in $\mathcal{T}'$ such that $\mathcal{T}' \models_{X'} \phi$.

In particular (taking $\phi = \bot$), if every finite $\mathcal{T}' \subseteq \mathcal{T}$ with free variables from finite $X' \subseteq X$ is satisfiable (over X'), then $\mathcal{T}$ is satisfiable (over X).

As a first application, we can show that Example 5.16 captures a general phenomenon of first-order logic: the inability to tell infinite cardinalities apart. Indeed, over an arbitrary signature $\mathcal{A}$, while it is not necessarily the case that *all* infinite structures are elementarily equivalent, the following guarantees that there are plenty of infinite structures which are:

Theorem 5.18. Let $\mathcal{T}$ be a (closed) first-order $\mathcal{A}$-theory which has models of cardinality $\geq n$ for each $n \in \mathbb{N}$, i.e.,

(i) either $\mathcal{T}$ has at least one infinite model,
(ii) or $\mathcal{T}$ has arbitrarily large finite models.

Then $\mathcal{T}$ has models of cardinality $\geq |X|$ for every set X.

Proof. WLOG we may assume that X is a set of variables (by Convention 3.23; otherwise, enlarge $\mathcal{V}$ to include X). Consider the open $\mathcal{A}$-theory

$$\mathcal{T}' := \mathcal{T} \cup \{\neg(x = y) \mid x \neq y \in X\}$$

with free variables from X. A model of it consists of an $\mathcal{A}$-structure $\mathcal{M}$ together with a variable assignment $\alpha : X \to M$ such that $\mathcal{M} \models_\alpha \mathcal{T}'$, which means $\mathcal{M} \models \mathcal{T}$ (since $\mathcal{T}$ has no free variables), and also $\mathcal{M} \models_\alpha \neg(x = y)$ for each $x \neq y \in X$, which means exactly that $\alpha : X \to M$ is injective. So $\mathcal{T}'$ is satisfiable iff $\mathcal{M}$ has a model of cardinality $\geq |X|$. By compactness, it suffices to show that for every finite $X' \subseteq X$ and $\mathcal{T}'' \subseteq \mathcal{T}'$ with free variables from X', there is a model $\mathcal{M}' \models_{\alpha'} \mathcal{T}''$ where $\alpha' : X' \to M'$. Indeed, since $|X'| \in \mathbb{N}$, we may let $\mathcal{M}'$ be a model of cardinality $\geq |X'|$, and $\alpha' : X' \to M'$ be an injection, so that $\mathcal{M}' \models_{\alpha'} \mathcal{T}''$ for the same reason as before. □

Example 5.19. The following classes $\mathcal{K}$ of structures are *not* axiomatizable in first-order logic:

- all finite sets
- all finite fields
- all finite abelian groups
- all finite posets
- all finite graphs
- ...

Indeed, since there are arbitrarily large finite structures of each of these types, any theory satisfied by all of them must also be satisfied by some infinite structure, by the preceding Theorem.

Example 5.20. For any infinite $\mathcal{A}$-structure $\mathcal{M}$, the class of structures isomorphic to $\mathcal{M}$ is *not* first-order axiomatizable, since any theory satisfied by $\mathcal{M}$ must also be satisfied by a structure of cardinality $> |M|$ (e.g., $\geq |\mathcal{P}(M)|$, which means $> |M|$ by Cantor's theorem).

In particular, this says that for infinite $\mathcal{M}$, its complete theory $\mathrm{Th}(\mathcal{M})$ has arbitrarily large models, i.e., there are arbitrarily large structures $\mathcal{N}$ elementarily equivalent to $\mathcal{M}$. Recall that (by Theorem 5.25 below) this means $\mathcal{M}, \mathcal{N}$ both elementarily embed into a common third structure. We can strengthen this conclusion, by tweaking the proof of Theorem 5.18 in the case that (i) holds:

Theorem 5.21 (upward Löwenheim–Skolem[12])**.** Let $\mathcal{M}$ be an infinite $\mathcal{A}$-structure. Then for any set X, there is an elementary embedding $\mathcal{M} \to \mathcal{N}$ into an $\mathcal{A}$-structure of cardinality $|N| \geq |X|$.

The proof uses the following device: the **elementary diagram** of an arbitrary $\mathcal{A}$-structure $\mathcal{M}$ is the complete theory

$$\mathrm{Th}_M(\mathcal{M}, \mathrm{id}_M),$$

[12]The full Löwenheim–Skolem theorem says that there is a model of exactly any infinite cardinality $\geq |\mathcal{A}|$, which admits an elementary embedding to or from $\mathcal{M}$ (depending on how its cardinality compares with that of $\mathcal{M}$). The proof of the "downward" part is by using an inductive procedure, similar to that in the proof of the completeness theorem, to show that every structure has a small substructure containing "enough" witnesses for all existentials.

where we regard each element $a \in M$ as a *variable*, which names itself under the identity variable assignment id_M. A model of $\mathrm{Th}_M(\mathcal{M}, \mathrm{id}_M)$ consists of an $\mathcal{A}$-structure $\mathcal{N}$ together with a "variable assignment" $h : M \to N$ such that

$$\mathcal{M} \models_{\mathrm{id}_M} \phi \iff \phi \in \mathrm{Th}_M(\mathcal{M}, \mathrm{id}_M) \implies \mathcal{N} \models_h \phi.$$

This implies that for any other variable assignment $\alpha : X \to M$, treating α instead as a variable *substitution* $\alpha : X \to M \subseteq \mathcal{L}^M_{\mathrm{term}}(\mathcal{A})$, we have

$$\begin{aligned} \mathcal{M} \models_\alpha \phi &\iff \mathcal{M} \models_{\alpha^{\mathcal{M}}(\mathrm{id}_M)} \phi \\ &\iff \mathcal{M} \models_{\mathrm{id}_M} \phi[\alpha] \qquad \text{by Lemma 4.41} \\ &\implies \mathcal{N} \models_h \phi[\alpha] \\ &\iff \mathcal{N} \models_{\alpha^{\mathcal{N}}(h) = h \circ \alpha} \phi \quad \text{by Lemma 4.41} \end{aligned} \tag{5.22}$$

(possibly after replacing ϕ with an α-equivalent copy to make $\phi[\alpha]$ safe). In other words, we have

$$\mathcal{N} \models_h \mathrm{Th}_M(\mathcal{M}, \mathrm{id}_M) \iff h : \mathcal{M} \to \mathcal{N} \text{ is an elementary embedding.}$$

Proof of Theorem 5.21. WLOG we may assume that $M \cap X = \varnothing$. Consider the open theory

$$\mathcal{T}' := \mathrm{Th}_M(\mathcal{M}, \mathrm{id}_M) \cup \{\neg(x = y) \mid x \neq y \in X\}$$

with free variables from $M \sqcup X$. A model consists of a structure $\mathcal{N}$ together with a variable assignment $\alpha : M \sqcup X \to N$ such that $\mathcal{N} \models_\alpha \mathcal{T}'$, i.e., $\mathcal{N} \models_{\alpha|M} \mathrm{Th}_M(\mathcal{M}, \mathrm{id}_M)$, which means $\alpha|M : \mathcal{M} \to \mathcal{N}$ is an elementary embedding, while $\alpha|X : X \hookrightarrow N$ is an injection as in the proof of Theorem 5.18. So by compactness, it suffices to show that for every finite $X' \subseteq X$, $\mathcal{T}' \cap \mathcal{L}^{M \sqcup X'}_{\mathrm{form}}(\mathcal{A})$ is satisfiable over $M \sqcup X'$. Indeed, since M is infinite, let $\alpha' : M \sqcup X' \to M$ be such that $\alpha'|M = \mathrm{id}_M : \mathcal{M} \to \mathcal{M}$ which is clearly an elementary embedding, while $\alpha'|X' : X' \hookrightarrow M$ is an injection; then $\mathcal{M} \models_{\alpha'} \mathcal{T}' \cap \mathcal{L}^{M \sqcup X'}_{\mathrm{form}}(\mathcal{A})$ for the same reason as before. □

Example 5.23. Consider $\mathbb{Z}$ equipped with the usual $+, \cdot, \leq$. By upward Löwenheim–Skolem, there is an elementary embedding $h : \mathbb{Z} \to \mathbb{Z}'$ into some uncountable $\{+, \cdot, \leq\}$-structure $\mathbb{Z}'$. Since $\mathbb{Z}$ is countable, this means that $\mathbb{Z}'$ must contain elements outside $\mathrm{im}(h)$. Since elementary embeddings are injective (see Exercise 2.64(a)), this means we may regard $\mathbb{Z}'$ as an "extension" of $\mathbb{Z}$ (or rather, $\mathrm{im}(h)$). Since $\mathbb{Z}$ is totally ordered via $\leq$, i.e., $\mathbb{Z} \models \mathcal{T}_{\mathsf{toset}}$, so must be $\mathbb{Z}'$. Since for each $n \in \mathbb{Z}$,

$$\mathbb{Z} \models_{x \mapsto n, y \mapsto n+1} (x \leq y) \wedge \forall z\, ((z \leq x) \vee (y \leq z)),$$

we get

$$\mathbb{Z}' \models_{x \mapsto h(n), y \mapsto h(h+1)} (x \leq y) \wedge \forall z\, ((z \leq x) \vee (y \leq z)),$$

i.e., the elements $\ldots, h(-1), h(0), h(1), h(2), \ldots$ form a consecutive range in the middle of $\mathbb{Z}'$, with no new elements in between. So the new elements in $\mathbb{Z}' \setminus \mathrm{im}(h)$ must be either $> h(n)$ for every $n \in \mathbb{Z}$, i.e., "positive infinite", or $< h(n)$ for every $n \in \mathbb{Z}$, i.e., "negative infinite". Note that since

$$\mathbb{Z} \models_{z \mapsto 0} \forall x\, \exists y\, (x + y = z),$$

$\mathbb{Z}'$ must have as many "positive infinite" elements as "negative infinite" elements. Similarly, $\mathbb{Z}'$ must obey all other laws of arithmetic that hold for the usual integers $\mathbb{Z}$.

This study of "infinite numbers" via first-order logic is the beginning of an area known as **nonstandard analysis**; elementary extensions of $\mathbb{Z}$ such as $\mathbb{Z}'$ are known as **hyperintegers**.

Remark 5.24. Can we give an explicit example of some such non-surjective elementary embedding $h : \mathbb{Z} \to \mathbb{Z}'$, i.e., an explicit system of "hyperintegers"? The proof of the upward Löwenheim–Skolem theorem uses the compactness theorem, which uses the completeness theorem, whose proof in Section 5 depends on a transfinite enumeration of all formulas, which depends on the Axiom of Choice (see Aside A, Section 6). In fact, a theorem of Tennenbaum (1959) says that there is *no* way to "computably" describe a system of hyperintegers!

The technique of elementary diagrams, which uses a theory to describe a *homomorphism* rather than just a structure, is quite powerful, and can be adapted to many other situations:

Theorem 5.25 (HW11)**.** Two $\mathcal{A}$-structures $\mathcal{M},\mathcal{N}$ are elementarily equivalent iff there is a third $\mathcal{A}$-structure $\mathcal{U}$ and elementary embeddings $\mathcal{M} \to \mathcal{U} \leftarrow \mathcal{N}$.

Theorem 5.26. Let $\mathcal{T}$ be an $\mathcal{A}$-theory, $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ be an $\mathcal{A}$-formula. The following are equivalent:

(i) there is a positive-existential $\psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ such that $\mathcal{T} \models_X \phi \leftrightarrow \psi$;
(ii) every homomorphism $h : \mathcal{M} \to \mathcal{N}$ between models of $\mathcal{T}$ preserves the interpretation of ϕ.

Proof. (i)$\implies$(ii) is by Proposition 2.49, which says that every homomorphism $h : \mathcal{M} \to \mathcal{N}$ preserves the interpretation of ψ, hence also that of ϕ if $\mathcal{M},\mathcal{N} \models \mathcal{T}$ since then ϕ,ψ are equivalent in $\mathcal{M},\mathcal{N}$.

For the converse, we define the **positive-existential diagram** of $\mathcal{M}$ to be

$$\text{Th}^{\exists+}_M(\mathcal{M}, \text{id}_M) := \{\text{positive-existential } \phi \in \mathcal{L}^M_{\text{form}}(\mathcal{A}) \mid \mathcal{M} \models_{\text{id}_M} \phi\}.$$

Exactly as for the elementary diagram (5.22), a model of $\text{Th}^{\exists+}_M(\mathcal{M}, \text{id}_M)$ consists of a structure $\mathcal{N}$ together with a "variable assignment" $h : M \to N$ which preserves the interpretation of all positive-existential formulas, which means exactly that h is a homomorphism $\mathcal{M} \to \mathcal{N}$ ($\Longleftarrow$ is by Proposition 2.49; $\Longrightarrow$ is by considering preservation of atomic formulas).

Lemma 5.27. For a fixed $\mathcal{A}$-structure $\mathcal{M}$ and variable assignment $\alpha : X \to M$ such that $\mathcal{M} \models_\alpha \phi$, the following are equivalent:

(i) there is a positive-existential $\psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ such that $\mathcal{M} \models_\alpha \psi$ and $\mathcal{T} \models_X \psi \to \phi$;
(ii) every homomorphism $h : \mathcal{M} \to \mathcal{N}$ into a model $\mathcal{N} \models \mathcal{T}$ preserves the interpretation of ϕ under α, i.e., obeys $\mathcal{N} \models_{h\circ\alpha} \phi$.

Proof. (i) $\implies$ (ii) is again by Proposition 2.49, which says that $\mathcal{N} \models_{h\circ\alpha} \psi$, whence $\mathcal{N} \models_{h\circ\alpha} \phi$ since $\mathcal{N} \models \mathcal{T} \models_X \psi \to \phi$.

Conversely, suppose (ii) holds; we prove (i). By the aforementioned connection between the positive-existential diagram and homomorphisms, (ii) can be restated as

$$\begin{aligned} \mathcal{N} \models_h \mathcal{T} \cup \text{Th}^{\exists+}_M(\mathcal{M}, \text{id}_M) &\implies \mathcal{N} \models_{h\circ\alpha} \phi \\ &\iff \mathcal{N} \models_h \phi[\alpha] \quad \text{by Lemma 4.41 as in (5.22)}, \end{aligned}$$

i.e.,

$$\mathcal{T} \cup \text{Th}^{\exists+}_M(\mathcal{M}, \text{id}_M) \models_M \phi[\alpha].$$

Thus by compactness, there is a finite $M' = \{a_1, \dots, a_n\} \subseteq M$ (with the a_i's distinct) containing all free variables from $\mathcal{T}' \subseteq \text{Th}^{\exists+}_M(\mathcal{M}, \text{id}_M)$ and $\phi[\alpha]$ such that

$$(*) \qquad \mathcal{T} \cup \mathcal{T}' \models_{M'} \phi[\alpha].$$

Reversing the above steps, this means that for every $\mathcal{N} \models \mathcal{T}$ and $h' : M' \to N$, we have

$$\mathcal{N} \models_{h'} \mathcal{T}' \implies \mathcal{N} \models_{h'} \phi[\alpha] \iff \mathcal{N} \models_{h'\circ\alpha|\text{FV}(\phi)} \phi \quad \text{by Lemma 4.41};$$

note that for $x \in \text{FV}(\phi)$, by Lemma 3.7, $\alpha(x) \in \bigcup_{y\in\text{FV}(\phi)} \text{FV}(\alpha(y)) = \text{FV}(\phi[\alpha]) \subseteq M'$, so that $h' \circ \alpha|\text{FV}(\phi) : \text{FV}(\phi) \to N$ is well-defined. Letting $\beta := h' \circ \alpha|\text{FV}(\phi)$, the above is equivalent to saying that for every $\beta : \text{FV}(\phi) \to N$,

$$(\exists h' : M' \to N \text{ s.t. } \beta = h' \circ \alpha|\text{FV}(\phi) \text{ and } \mathcal{N} \models_{h'} \mathcal{T}') \implies \mathcal{N} \models_\beta \phi;$$

now letting $M' := \{a_1, \dots, a_n\}$ (without repetitions), denoting $h'(a_1), \dots, h'(a_n)$ by new variables $y_1, \dots, y_n$ not appearing anywhere, and for each $x \in \text{FV}(\phi)$, letting $\alpha(x) = a_{i_x} \in M'$, this says

$$\mathcal{N} \models_\beta \exists y_1 \cdots \exists y_n \, (\textstyle\bigwedge_{x\in\text{FV}(\phi)}(x = y_{i_x}) \wedge \bigwedge \mathcal{T}'[a_i \mapsto y_i]) \implies \mathcal{N} \models_\beta \phi.$$

So letting ψ be this formula on the LHS, we have $\mathcal{T} \models_X \psi \to \phi$, and $\mathcal{M} \models_\alpha \psi$ under $y_i \mapsto a_i$. $\square$

Now to prove (ii) $\implies$ (i) in the Theorem: (ii) is equivalent by the Lemma to

$$\forall \mathcal{M},\ (\mathcal{M} \models \mathcal{T} \text{ and } \mathcal{M} \models_\alpha \phi \implies \exists \text{ pos.-exist. } \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A}) \text{ s.t. } \mathcal{T} \models_X \psi \to \phi \text{ and } \mathcal{M} \models_\alpha \psi),$$

i.e.,

$$\mathcal{T} \cup \{\phi\} \cup \{\neg\psi \mid \text{pos.-exist. } \psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A}) \text{ s.t. } \mathcal{T} \models_X \psi \to \phi\} \models_X \bot.$$

By compactness, some subset of this theory including only finitely many of the formulas $\neg\psi$ is already unsatisfiable, i.e., there is a finite set $\Psi \subseteq \mathcal{L}^X_{\text{form}}(\mathcal{A})$ of positive-existential formulas each of which together with $\mathcal{T}$ implies ϕ over X, such that

$$\begin{aligned} \mathcal{M} \models \mathcal{T} \text{ and } \mathcal{M} \models_\alpha \phi &\implies \exists \psi \in \Psi \text{ s.t. } \mathcal{M} \models_\alpha \psi \\ &\iff \mathcal{M} \models_\alpha \bigvee \Psi, \end{aligned}$$

which means that $\mathcal{T} \models_X \phi \to \bigvee \Psi$. On the other hand, $\mathcal{T} \models_X \bigvee \Psi \to \phi$, since $\mathcal{T} \models_X \psi \to \phi$ for each $\psi \in \Psi$. So $\bigvee \Psi$ works as ψ in (i). □

Exercise 5.28 (HW11)**.** The **quantifier-free diagram** $\text{Th}^{\text{qf}}_M(\mathcal{M}, \text{id}_M)$ of a structure $\mathcal{M}$ is the set of all quantifier-free formulas satisfied by $\mathcal{M}$ under the identity variable assignment $\text{id}_M : M \to M$. Prove that a model of $\text{Th}^{\text{qf}}_M(\mathcal{M}, \text{id}_M)$ is the same thing as a structure $\mathcal{N}$ together with a homomorphism $h : \mathcal{M} \to \mathcal{N}$ which is an isomorphism with its image substructure.

Theorem 5.29 (HW11)**.** Let $\mathcal{T}$ be an $\mathcal{A}$-theory, $\phi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$. The following are equivalent:

(i) there is a finite conjunction of universal formulas $\psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ such that $\mathcal{T} \models_X \phi \leftrightarrow \psi$;
(ii) for every $\mathcal{N} \models \mathcal{T}$ and substructure $\mathcal{M} \models \mathcal{T}$ which is also a model of $\mathcal{T}$, for every $\alpha : X \to M$, if $\mathcal{N} \models_\alpha \phi$, then $\mathcal{M} \models_\alpha \phi$.

Proof. (i) $\implies$ (ii) is by HW6. For the converse, one first proves the following:

Lemma 5.30. For a fixed $\mathcal{A}$-structure $\mathcal{M}$ and variable assignment $\alpha : X \to M$ such that $\mathcal{M} \models_\alpha \phi$, the following are equivalent:

(i) there is an existential $\psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ such that $\mathcal{M} \models_\alpha \psi$ and $\mathcal{T} \models_X \psi \to \phi$;
(ii) every homomorphism $h : \mathcal{M} \to \mathcal{N}$ into a model $\mathcal{N} \models \mathcal{T}$ and which is an isomorphism with its image substructure obeys $\mathcal{N} \models_{h\circ\alpha} \phi$.

Proof. (i) $\implies$ (ii) is again by HW6: from $\mathcal{M} \models_\alpha \psi$, we get $\text{im}(h) \models_{h\circ\alpha} \psi$ since $h : \mathcal{M} \to \text{im}(h)$ is an isomorphism, whence $\mathcal{N} \models_{h\circ\alpha} \psi$ by HW6 since $\neg\psi$ is equivalent to a universal formula, whence $\mathcal{N} \models_{h\circ\alpha} \phi$ since $\mathcal{T} \models_X \psi \to \phi$. The proof of (ii) $\implies$ (i) is identical to that of Lemma 5.27, with the positive-existential diagram replaced by the quantifier-free diagram. □

Now to prove (ii) $\implies$ (i) in the Theorem: from (ii), we get that for every homomorphism $h : \mathcal{M} \to \mathcal{N}$ between models of $\mathcal{T}$ which is an isomorphism with its image substructure, for every $\alpha : X \to M$, if $\mathcal{M} \models_\alpha \neg\phi$, then $\text{im}(h) \models_{h\circ\alpha} \neg\phi$ since $h : \mathcal{M} \to \text{im}(h)$ is an isomorphism, whence $\mathcal{N} \models_{h\circ\alpha} \neg\phi$ by (ii). Imitating the proof of Theorem 5.26 using the above Lemma, we get that $\neg\phi$ is $\mathcal{T}$-equivalent to a finite disjunction of existential formulas, hence ϕ is $\mathcal{T}$-equivalent to a finite conjunction of universal formulas. □

From Lemma 5.30, we may also deduce the following interesting consequence:

Corollary 5.31. Let $\mathcal{T}$ be an $\mathcal{A}$-theory. Then an $\mathcal{A}$-structure $\mathcal{M}$ is isomorphic to a substructure of a model of $\mathcal{T}$ iff it satisfies all universal consequences of $\mathcal{T}$.

Proof. $\implies$ is again by HW6. Conversely, if $\mathcal{M}$ is not isomorphic to any substructure of a model of $\mathcal{T}$, then every homomorphism $h : \mathcal{M} \to \mathcal{N}$ with $\mathcal{N} \models \mathcal{T}$ and $h : \mathcal{M} \cong \text{im}(h)$ (vacuously) obeys $\mathcal{N} \models \bot$, whence by Lemma 5.30, there is an existential $\psi \in \mathcal{L}^X_{\text{form}}(\mathcal{A})$ such that $\mathcal{M} \models_\alpha \psi$ and $\mathcal{T} \models \psi \to \bot$, whence $\neg\psi$ is equivalent to a universal consequence of $\mathcal{T}$ which is false in $\mathcal{M}$. □