The Disjunctive Normal Form Theorem

Intermediate Logic

September 13, 2011

The disjunctive normal form (DNF) theorem is among the most central facts of truth-functional logic. Its centrality is two-fold:

1. It is a powerful tool that leads to, among other things, an efficient decision procedure for truth-functional logic.

2. There are several different approaches, based on the several methods of truth-functional evaluation, that one can take to proving it.

The purpose of these notes is to present a "constructive" proof of the DNF theorem, from which one can extract a universal procedure for transforming in a step-by step manner any truth-functional formula with connectives $\wedge$, $\vee$, $\neg$, $\supset$, and $\equiv$ into an equivalent formula in DNF.

We begin with the basic definitions:

- A *literal* is a sentence letter or the negation of a sentence letter.

- A formula is in *disjunctive normal form* if it is a disjunction of conjunctions of literals. In this definition, we allow for the "dummy cases" of a conjunction with only one conjunct (thus not including the symbol $\wedge$) and of a disjunction with only one disjunct (thus not including the symbol $\vee$.) It is perhaps more natural to think about what cannot occur in a DNF formula: The connectives $\supset$, $\equiv$ cannot occur at all, the connective $\neg$ cannot occur except in a literal, the connective $\wedge$ can only govern literals. (The appearance of the connective $\vee$ is also resticted in that it can only govern conjunctions of literals, although this restriction is automatically satisfied whenever the others are.) To properly understand the restriction on conjunctions (and on disjunctions) recall our convention of dropping parentheses in sequences of conjunctions, so that the conjunctions in the formula $p \wedge \neg q \wedge r \wedge s$ each count as governing literals. (One easily sees that the characterization of DNF formulas as disjunctions of conjunctions of literals amounts to the same thing as imposing these restrictions on how connectives can appear.)

- The disjuncts of a DNF formula are called its *clauses*.

Our proof uses the fact that every formula is equivalent to a Boolean formula (i.e., one with only the connectives $\wedge$, $\vee$, and $\neg$.) It also uses four other facts of equivalence: the equivalence of $\neg\neg S$ and S, the equivalence of $(S_1 \vee S_2 \vee \ldots \vee S_n) \wedge T$ and $(S_1 \wedge T \vee S_2 \wedge T \vee \ldots \vee S_n \wedge T)$, the equivalence of $\neg(S_1 \vee S_2 \vee \ldots \vee S_n)$ and $\neg S_1 \wedge \neg S_2 \wedge \ldots \wedge \neg S_n$, and the equivalence of $\neg(S_1 \wedge S_2 \wedge \ldots \wedge S_n)$ and $\neg S_1 \vee \neg S_2 \vee \ldots \wedge \vee S_n$. All of these are subject to simple inductive proofs. For the last four facts, we are only interested in "one direction" of the equivalence, namely the transformation from the left to the right. We call the first two of these transformation rules "double negation elimination" and "distribution (of conjunction over disjunction)." We call each of the last two "DeMorgan's inward." Notice that the interchange theorem allows us to perform these transformations not only *to* a formula but also *within* a formula.

The basic idea of our proof is simple. We claim that any formula, once it is rewritten without the connectives $\supset$ and $\equiv$, can be transformed into DNF by successive applications of these four rules. The constructive nature of the proof results from the fact that, not only will we be able to see *that* the theorem is true and *why* the theorem is true, and not only will we be able to find an actual DNF equivalent to any truth-functional formula, but one will be able to trace the equivalence through a sequence of formulas each of which is *obviously* equivalent to its predecessor.

Our claim is not immediately obviously true, though, because a ballistic application of DeMorgan's inward and distribution can lead one into horrible unwieldiness. Consider the following simple formula: $\neg((p \vee q \vee r) \wedge (q \vee \neg r \vee s))$. Below is the start of a transformation of this formula according to the basic idea described above.

$$\neg((p \vee q \vee r) \wedge (q \vee \neg r \vee s))$$
$$\neg((p \wedge (q \vee \neg r \vee s)) \vee (q \wedge (q \vee \neg r \vee s)) \vee (r \wedge (q \vee \neg r \vee s)))$$
$$\neg((p \wedge q) \vee (p \wedge \neg r) \vee (p \wedge s) \vee (q \wedge q) \vee (q \wedge \neg r) \vee (q \wedge s) \vee (r \wedge q) \vee (r \wedge \neg r) \vee (r \wedge s))$$
$$\neg(p \wedge q) \wedge \neg(p \wedge \neg r) \wedge \neg(p \wedge s) \wedge \neg(q \wedge q) \wedge \neg(q \wedge \neg r) \wedge \neg(q \wedge s) \wedge \neg(r \wedge q) \wedge \neg(r \wedge \neg r) \wedge \neg(r \wedge s)$$
$$(\neg p \vee \neg q) \wedge (\neg p \vee \neg\neg r) \wedge \ldots$$

This is getting awfully complex, and, what's worse, at this point the horror is still mounting: one can see that several more opportunities for distribution are arising in the last step. A natural feeling at such times is that one is making no progress and perhaps only making things worse. This intuition can be made precise by defining a *bad connective* as a connective that violates one of the constraints of DNF. Our basic idea, then, is to apply successive transformations until all the bad connectives vanish. The example above challenges the claim that this is always possible by showing that bad connectives may actually accumulate as one proceeds. How does one know how to apply the rules in a way that steers safe of an endless cycle of distributions? How does one even know that there aren't some pathological formulas that admit no safe steering whatever?

Our solution to this conundrum proceeds by induction on the number of bad connectives in a formula. Since we know that $\supset$ and $\equiv$ are eliminable, let us consider a formula Φ without these two connectives.

Suppose, first, that Φ has only one bad connective. One sees immediately that the connective must be either $\wedge$ or $\neg$, for the only other connective is $\vee$ and if Φ were of the form (DNF) $\vee$ (DNF) (i.e., a disjunction of two DNF sub-formulas), then Φ would itself be in DNF and would therefore have no bad connectives.

There are three cases to consider:

1	Φ is $\Psi \wedge X$	where Ψ and X are DNF formulas
2	Φ is $\neg\Psi$	where Ψ is a one-clause DNF formula
3	Φ is $\neg\Psi$	where Ψ is a multi-clause DNF formula

In **case 1**, if both Ψ and X are one-clause DNF formulas, then Φ already is in DNF and there is nothing to do. Otherwise, if only one sub-formula is a multi-clause DNF formula, then one application of distribution results in a DNF equivalent to Φ, and if both sub-formulas are multi-clause, then distributing each clause of Ψ over X after distributing X over Ψ results in a DNF equivalent to Φ.

(As a bonus **case 1***, notice that any formula all of whose bad connectives are conjunction symbols can be transformed by successive applications of distribution to a DNF equivalent. For any such formula has the form $(DNF) \wedge (DNF) \wedge \ldots \wedge (DNF)$. (Just reason inductively as follows: We just proved that it's possible if there is only one bad conjunction symbol. Now, suppose it's possible when there are n bad conjunction symbols. Then, given a formula Φ with $n+1$ bad conjunction symbols, reduce the sub-formula of Φ containing the first n bad conjunction symbols to DNF, and apply the interchange theorem to get an equivalent to Φ whose only bad connective is a single $\wedge$ (the $n+1$st). Apply distribution rules to this new formula as in case 1 to get a DNF equivalent to Φ.)

In **case 2**, a single application of DeMorgan's inward yields a disjunction Ψ of sentence letters, negated sentence letters, and doubly negated sentence letters. Apply double negation elimination to all the doubly negated sentence letters and interchange the resulting sentence letters back into Ψ. The result is a DNF equivalent to Φ.

In **case 3**, Φ has the form $\neg(C_1 \vee C_2 \vee \ldots \vee C_n)$, where the C_i are conjunctions of literals. First, apply DeMorgan's inward to get $\neg C_1 \wedge \neg C_2 \wedge \ldots \wedge \neg C_n$. Then, for each C_i that contains a conjunction symbol (i.e., for each one that is not a single literal), replace $\neg C_i$ with its DNF equivalent attainable as in case 2. The result is a formula whose only bad connectives are conjunction symbols, which can be transformed to a DNF equivalent as in case 1*.

Thus we see that every formula with only one bad connective can be transformed to a DNF equivalent. For the induction step, assume that all formulas with n or fewer bad connectives can be thus transformed, and suppose now that Φ has $n+1$ bad connectives. There are again three possibilities:

1	Φ is $\Psi \vee X$	where Ψ and X each have n or fewer bad connectives
2	Φ is $\Psi \wedge X$	where Ψ and X each have n or fewer bad connectives
3	Φ is $\neg\Psi$	where Ψ has n or fewer bad connectives

In each case, apply the induction hypothesis to find DNF equivalents Ψ_{DNF} and X_{DNF} to Ψ and X, and apply the interchange theorem to create an equivalent to Φ (either $(\Psi_{DNF}) \vee (X_{DNF})$, $(\Psi_{DNF}) \wedge (X_{DNF})$, or $\neg(\Psi_{DNF})$). The first possibility is already in DNF. The second and third possibilities each contain only one bad connective and are thus subject to the treatment in the base step.

From the inductive nature of this proof one can extract a procedure for actually transforming formulas to their DNF equivalents. One would attack each formula "from the inside out," applying transformation rules typically as close to the level of literals as possible. Let us stress that this is theoretically advantageous for letting one keep track of one's progress towards DNF, but it is not prudent strategy for actual reductions. In fact, the best practical approach is exactly the opposite: work one's way in from the main connective.

Consider again the formula: $\neg((p \vee q \vee r) \wedge (q \vee \neg r \vee s))$. Below is the full transformation of this formula according to the best practical approach:

$$\begin{gathered}
\neg((p \vee q \vee r) \wedge (q \vee \neg r \vee s) \\
\neg(p \vee q \vee r) \vee \neg(q \vee \neg r \vee s) \\
(\neg p \wedge \neg q \wedge \neg r) \vee \neg(q \vee \neg r \vee s) \\
(\neg p \wedge \neg q \wedge \neg r) \vee (\neg q \wedge r \wedge \neg s)
\end{gathered}$$

This example suggests that it is efficient to apply transformations to the highest ranking connectives possible. A little practice should convince you that this suggestion is accurate, and modest reflection on the above proof of the DNF theorem should convince you that the process will eventually terminate. However, all proofs that I know of the DNF theorem that make perspicuous the "outside in" advantage are comparably less easy to follow.