

Antimirov's Proof about Pders

These are some rough notes about the result by Antimirov establishing a bound on the number of regular expressions in a partial derivative. From this bound on the number of partial derivatives one can easily construct an NFA for a regular expression, but one can also derive a bound on the size of the partial derivatives. This is what we do first. We start with the following definitions:

- $pder\ c\ r$ — partial derivative according to a character; this can be defined inductively as follows:

$$\begin{aligned}
 pder\ c\ (\mathbf{0}) &\stackrel{\text{def}}{=} \emptyset \\
 pder\ c\ (\mathbf{1}) &\stackrel{\text{def}}{=} \emptyset \\
 pder\ c\ (d) &\stackrel{\text{def}}{=} \text{if } c = d \text{ then } \{\mathbf{1}\} \text{ else } \emptyset \\
 pder\ c\ (r_1 + r_2) &\stackrel{\text{def}}{=} pder\ c\ r_1 \cup pder\ c\ r_2 \\
 pder\ c\ (r_1 \cdot r_2) &\stackrel{\text{def}}{=} \text{if } nullable(r_1) \\
 &\quad \text{then } \Pi(pder\ c\ r_1)\ r_2 \cup pder\ c\ r_2 \\
 &\quad \text{else } \Pi(pder\ c\ r_1)\ r_2 \\
 pder\ c\ (r^*) &\stackrel{\text{def}}{=} \Pi(pder\ c\ r)\ (r^*)
 \end{aligned}$$

- $pder^+\ c\ rs$ — partial derivatives for a set regular expressions rs
- $pders\ s\ r$ — partial derivative of a regular expression according to a string
- $Pders\ A\ r \stackrel{\text{def}}{=} \bigcup_{s \in A} pders\ s\ r$ — partial derivatives according to a language (a set of strings)
- $|rs|$ is the size of a set of regular expressions rs , or the number of elements in the set (also known as the cardinality of this set)
- $\prod rs\ r \stackrel{\text{def}}{=} \{r_1 \cdot r \mid r_1 \in rs\}$ — this is some convenience when writing a set of sequence regular expressions. It essentially “appends” the regular expression r to all regular expressions in the set rs . As a result

one can write the sequence case for partial derivatives (see above) more conveniently as

$$pder\ c\ (r_1 \cdot r_2) \stackrel{\text{def}}{=} \begin{cases} \prod (pder\ c\ r_1) r_2 \cup pder\ c\ r_2 & \text{provided } r_1 \text{ is nullable} \\ \prod (pder\ c\ r_1) r_2 & \text{otherwise} \end{cases}$$

- $Psufs$ is the set of all non-empty suffixes of s defined as

$$PSufs \stackrel{\text{def}}{=} \{v. v \neq [] \wedge \exists u. u @ v = s\}$$

So for the string abc the non-empty suffixes are c , bc and abc . Also we have that $Psuf(s @ [c]) = ((Psufs) @@ [c]) \cup \{[c]\}$. Here $@@$ means to concatenate $[c]$ to the end of all strings in $Psufs$; in this equation we also need to add $\{[c]\}$ in order to make the equation to hold.

To state Antimirov's result we need the following definition of an *alphabetic width* of a regular expression defined as follows:

$$\begin{aligned} awidth(\mathbf{0}) & \stackrel{\text{def}}{=} 0 \\ awidth(\mathbf{1}) & \stackrel{\text{def}}{=} 0 \\ awidth(c) & \stackrel{\text{def}}{=} 1 \\ awidth(r_1 + r_2) & \stackrel{\text{def}}{=} awidth(r_1) + awidth(r_2) \\ awidth(r_1 \cdot r_2) & \stackrel{\text{def}}{=} awidth(r_1) + awidth(r_2) \\ awidth(r^*) & \stackrel{\text{def}}{=} awidth(r) \end{aligned}$$

This function counts how many characters are in a regular expression. Antimirov's result states

Theorem 1 $\forall A\ r. |Pders\ A\ r| \leq awidth(r) + 1$

Note this theorem holds for any set of strings A , for example for the set of all strings, which I will write as $UNIV$, and also for the set $\{s\}$ containing only a single string s . Therefore a simple corollary is

Corollary 2 $\forall s\ r. |pders\ s\ r| \leq awidth(r) + 1$

This property says that for every string s , the number of regular expressions in the derivative can never be bigger than $awidth(r) + 1$. Interestingly we do not show Thm 1 for all sets of strings A directly, but rather only for one particular set of strings which I call $UNIV_1$. It includes all strings except the empty string (remember $UNIV$ contains all strings).

Let's try to give below a comprehensible account of Antimirov's proof of Thm. 1—I distinctly remember that Antimirov's paper is great, but pretty incomprehensible for the first 20+ times one reads that paper. The proof starts with the following much weaker property about the size being finite:

Lemma 3 $\forall A r. (Pders A r) \text{ is a finite set.}$

This lemma is needed because some reasoning steps in Thm 1 only work for finite sets, not infinite sets. But let us skip over the proof of this property at first and let us assume we know already that the partial derivatives are always finite sets (this for example does not hold for unsimplified Brzozowski derivatives which can be infinite for some sets of strings).

There are some central lemmas about partial derivatives for \cdot and $_*$. One is the following

Lemma 4

$$Pders UNIV_1 (r_1 \cdot r_2) \subseteq (\prod (Pders UNIV_1 r_1) r_2) \cup Pders UNIV_1 r_2$$

Proof: *The proof is done via an induction for the following property*

$$pders s (r_1 \cdot r_2) \subseteq (\prod (pders s r_1) r_2) \cup Pders (PSufs) r_2$$

Note that this property uses $pders$ and $Pders$ together. The proof is done by “reverse” induction on s , meaning the cases to analyse are the empty string $[]$ and the case where a character is put at the end of the string s , namely $s@[c]$. The case $[]$ is trivial. In the other case we know by IH that

$$pders s (r_1 \cdot r_2) \subseteq (\prod (pders s r_1) r_2) \cup Pders (PSufs) r_2$$

holds for s . Then we have to show it holds for $s@[c]$

$$\begin{aligned}
& pders(s@[c])(r_1 \cdot r_2) \\
&= pder^+ c(pders s(r_1 \cdot r_2)) \\
&\subseteq pder^+ c(\prod(pders s r_1) r_2 \cup Pders(PSufs) r_2) \\
&\hspace{15em} \text{by IH} \\
&= (pder^+ c(\prod(pders s r_1) r_2)) \cup (pder^+ c(Pders(PSufs) r_2)) \\
&= (pder^+ c(\prod(pders s r_1) r_2)) \cup (Pders(PSuf(s@[c])) r_2) \\
&\subseteq (pder^+ c(\prod(pders s r_1) r_2)) \cup (pder c r_2) \cup (Pders(PSufs@@[c]) r_2) \\
&\subseteq \prod(pder^+ c(pders s r_1)) r_2 \cup (pder c r_2) \cup (Pders(PSufs@@[c]) r_2) \\
&= (\prod(pders(s@[c]) r_1) r_2) \cup (pder c r_2) \cup (Pders(PSufs@@[c]) r_2) \\
&\subseteq (\prod(pders(s@[c]) r_1) r_2) \cup (Pders(PSuf(s@[c])) r_2)
\end{aligned}$$

The lemma now follows because for an $s \in UNIV_1$ it holds that

$$\prod(pders s r_1) r_2 \subseteq \prod(Pders UNIV_1 r_1) r_2$$

and

$$Pders(PSufs) r_2 \subseteq Pders UNIV_1 r_2$$

The left-hand sides of the inclusions above are equal to $pders s(r_1 \cdot r_2)$ for a string $s \in UNIV_1$. \square

There is a similar lemma for the $*$ -regular expression, namely:

Lemma 5 $Pders UNIV_1(r^*) \subseteq \prod(Pders UNIV_1 r)(r^*)$

We omit the proof for the moment (similar to Lem 4). We also need the following property about the cardinality of \prod :

Lemma 6 $|\prod(Pders A r_1) r_2| \leq |Pders A r_1|$

We only need the \leq version, which essentially says there are as many sequences $r \cdot r_2$ as are elements in r . Now for the proof of Thm 1: The main induction in Antimirov's proof establishes that:¹

Lemma 7 $\forall r. |Pders UNIV_1 r| \leq awidth(r)$

¹Remember that it is always the hardest part in an induction proof to find the right property that is strong enough and of the right shape for the induction to go through.

Proof: This is proved by induction on r . The interesting cases are $r_1 + r_2$, $r_1 \cdot r_2$ and r^* . Let us start with the relatively simple case:

Case $r_1 + r_2$: By induction hypothesis we know

$$\begin{aligned} |Pders \text{ UNIV}_1 r_1| &\leq awidth(r_1) \\ |Pders \text{ UNIV}_1 r_2| &\leq awidth(r_2) \end{aligned}$$

In this case we can reason as follows

$$\begin{aligned} &|Pders \text{ UNIV}_1 (r_1 + r_2)| \\ &= |(Pders \text{ UNIV}_1 r_1) \cup (Pders \text{ UNIV}_1 r_2)| \\ &\leq |(Pders \text{ UNIV}_1 r_1)| + |(Pders \text{ UNIV}_1 r_2)| \quad (*) \\ &\leq awidth(r_1) + awidth(r_2) \\ &\stackrel{\text{def}}{=} awidth(r_1 + r_2) \end{aligned}$$

Note that $(*)$ is a step that only works if one knows that $|(Pders \text{ UNIV}_1 r_1)|$ and so on are finite. The next case is a bit more interesting:

Case $r_1 \cdot r_2$: We have the same induction hypothesis as in the case before.

$$\begin{aligned} &|Pders \text{ UNIV}_1 (r_1 \cdot r_2)| \\ &\leq |\prod (Pders \text{ UNIV}_1 r_1) r_2 \cup (Pders \text{ UNIV}_1 r_2)| \quad \text{by Lem 4} \\ &\leq |\prod (Pders \text{ UNIV}_1 r_1) r_2| + |(Pders \text{ UNIV}_1 r_2)| \\ &\leq |Pders \text{ UNIV}_1 r_1| + |Pders \text{ UNIV}_1 r_2| \quad \text{by Lem 6} \\ &\leq awidth(r_1) + awidth(r_2) \\ &\stackrel{\text{def}}{=} awidth(r_1 \cdot r_2) \end{aligned}$$

Case r^* : Again we have the same induction hypothesis as in the cases before.

$$\begin{aligned} &|Pders \text{ UNIV}_1 (r^*)| \\ &\leq |\prod (Pders \text{ UNIV}_1 r) (r^*)| \quad \text{by Lem 5} \\ &\leq |Pders \text{ UNIV}_1 r| \quad \text{by Lem 6} \\ &\leq awidth(r) \end{aligned}$$

□

From this lemma we can derive the next corollary which extends the property to $\text{UNIV} (= \text{UNIV}_1 \cup \{\square\})$:

Corollary 8 $\forall r. |Pders \text{ UNIV } r| \leq awidth(r) + 1$

Proof: *This can be proved as follows*

$$\begin{aligned}
& |Pders \text{ UNIV } r| \\
&= |Pders (\text{UNIV}_1 \cup \{\emptyset\}) r| \\
&= |(Pders \text{ UNIV}_1 r) \cup \{r\}| \\
&\leq |Pders \text{ UNIV}_1 r| + 1 && \text{by Lem 7} \\
&\leq awidth(r) + 1
\end{aligned}$$

□

From the last corollary, it is easy to infer Antimirov's Thm 1, because

$$Pders A r \subseteq Pders \text{ UNIV } r$$

for all sets A .

While I was earlier a bit dismissive above about the intelligibility of Antimirov's paper, you have to admit this proof is a work of beauty. It only gives a bound ($awidth$) for the number of regular expressions in the derivatives—this is important for constructing NFAs. Maybe it has not been important before, but I have never seen a result about the *size* of the partial derivatives.² However, a very crude bound, namely $(size(r)^2 + 1) \times (awidth(r) + 1)$, can be easily derived by using Antimirov's result. The definition we need for this is a function that collects subexpressions from which partial derivatives are built:

$$\begin{aligned}
subs(\mathbf{0}) &\stackrel{\text{def}}{=} \{\mathbf{0}\} \\
subs(\mathbf{1}) &\stackrel{\text{def}}{=} \{\mathbf{1}\} \\
subs(c) &\stackrel{\text{def}}{=} \{c, \mathbf{1}\} \\
subs(r_1 + r_2) &\stackrel{\text{def}}{=} \{r_1 + r_2\} \cup subs(r_1) \cup subs(r_2) \\
subs(r_1 \cdot r_2) &\stackrel{\text{def}}{=} \{r_1 \cdot r_2\} \cup (\prod subs(r_1) r_2) \cup subs(r_1) \cup subs(r_2) \\
subs(r^*) &\stackrel{\text{def}}{=} \{r^*\} \cup (\prod subs(r) r^*) \cup subs(r)
\end{aligned}$$

We can show that

Lemma 9 $pders s r \subseteq subs(r)$

This is a relatively simple induction on r . The point is that for every element in $subs$, the maximum size is given by

²Update: I have now seen a paper which proves this result as well.

Lemma 10 *If $r' \in \text{subs}(r)$ then $\text{size}(r') \leq 1 + \text{size}(r)^2$.*

Again the proof is a relatively simple induction on r . Stringing Antimirov's result and the lemma above together gives

Theorem 11 $\sum_{r' \in \text{pders } s r} \text{size}(r') \leq (\text{size}(r)^2 + 1) \times (\text{awidth}(r) + 1)$

Since *awidth* is always smaller than the *size* of a regular expression, one can also state the bound as follows:

$$\sum_{r' \in \text{pders } s r} \text{size}(r') \leq (\text{size}(r) + 1)^3$$

This, by the way, also holds for *Pders*, namely

$$\sum_{r' \in \text{Pders } A r} \text{size}(r') \leq (\text{size}(r) + 1)^3$$

for all r and A . If one is interested in the height of the partial derivatives, one can derive:

$$\forall r' \in \text{pders } s r. \text{height}(r') \leq \text{height}(r) + 1$$

meaning the height of the partial derivatives is never bigger than the height of the original regular expression (+1).

NFA Construction via Antimirov's Partial Derivatives

Let's finish these notes with the construction of an NFA for a regular expression using partial derivatives. As usual an automaton is a quintuple (Q, A, δ, q_0, F) where Q is the set of states of the automaton, A is the alphabet, q_0 is the starting state and F are the accepting states. For DFAs the δ is a (partial) function from state \times character to state. For NFAs it is a relation between state \times character \times state. The non-determinism can be seen by the following: consider three (distinct) states q_1 , q_2 and q_3 , then the relation can include (q_1, a, q_2) and (q_1, a, q_3) , which means there is a transition between q_1 and both q_2 and q_3 for the character a .

The Antimirov's NFA for a regular expression r is then given by the quintuple

$$(PD(r), A, \delta_{PD}, r, F)$$

where $PD(r)$ are all the partial derivatives according to all strings, that is

$$PD(r) \stackrel{\text{def}}{=} Pders \text{ UNIV } r$$

Because of the previous proof, we know that this set is finite. We also see that the states in Antimirov's NFA are "labelled" by single regular expressions. The starting state is labelled with the original regular expression r . The set of accepting states F is all states $r' \in F$ where r' is nullable. The relation δ_{PD} is given by

$$(r_1, c, r_2)$$

for every $r_1 \in PD(r)$ and $r_2 \in pder \ c \ r$. This is in general a "non-deterministic" relation because the set of partial derivatives often contains more than one element. A nice example of an NFA constructed via Antimirov's partial derivatives is given in [1] on Page 378.

The difficulty of course in this construction is to find the set of partial derivatives according to *all* strings. However, it seem a procedure that enumerates strings according to size suffices until no new derivative is found. There are various improvements that apply clever tricks on how to more efficiently discover this set.

References

- [1] L. Ilie and S. Yu, *Reducing NFAs by Invariant Equivalences*. In Theoretical Computer Science, Volume 306(1–3), Pages 373–390, 2003.
<https://core.ac.uk/download/pdf/82545723.pdf>