Every context-free grammar over a terminal alphabet of cardinality 1 generates a regular language.

Let us consider a context-free grammar G which, without loss of generality, does not have ε-productions besides, possibly, the production $S \to \varepsilon$.

We want to show that if the terminal alphabet of G is a singleton, then the language $L(G)$ generated by the grammar G is a regular language.

Given a word w, by $|w|$ we will denote the length of w.

Let us first recall the Pumping Lemma for context-free languages.

Lemma 1. [**Pumping Lemma**] *Given a context-free grammar G with terminal alphabet Σ, $\exists\, n>0$ such that $\forall\, z \in L(G)$, if $|z| \geq n$ then $\exists\, u, v, w, x, y \in \Sigma^*$, such that*
(1) $z = uvwxy$,
(2) $vx \neq \varepsilon$,
(3) $|vwx| \leq n$, *and*
(4) $\forall\, i \geq 0$, $uv^iwx^iy \in L(G)$.

Let us assume that the terminal alphabet of G is the set $\Sigma = \{a\}$ with cardinality 1. Since Σ has cardinality 1, commutativity holds, that is, $\forall\, u, v \in \Sigma^*$, $u\,v = v\,u$.

The following lemma easily follows from Lemma 1.

Lemma 2. [**Pumping Lemma for a Terminal Alphabet of Cardinality 1**] *Given a context-free grammar G with a terminal alphabet Σ of cardinality 1, $\exists\, n>0$ such that $\forall\, z \in L(G)$, if $|z| \geq n$ then $\exists\, p \geq 0, \exists\, q$, such that*
(1.1) $|z| = p+q$,
(2.1) $q>0$,
(3.1) $\exists\, m$, *with* $0 \leq m \leq p$, *such that* $0 < m+q \leq n$, *and*
(4.1) $\forall\, s \in \Sigma^*, \forall\, i \geq 0$, *if* $|s| = p + i\,q$ *then* $s \in L(G)$.

Proof. The final part of the statement of Lemma 1 can be rewritten as follows. By commutativity, we can absorb vx into v (note that v and x are both existentially quantified) and we get:

$\ldots \exists\, u, v, w, y \in \Sigma^*$, *such that*
$z = uvwy$,
$v \neq \varepsilon$,
$|vw| \leq n$, *and*
$\forall\, i \geq 0$, $uv^iwy \in L(G)$.

By commutativity, we can absorb uy into u (note that u and y are both existentially quantified) and we get:

$\ldots \exists\, u, v, w \in \Sigma^*$, *such that*
$z = uvw$,
$v \neq \varepsilon$,
$|vw| \leq n$, *and*
$\forall\, i \geq 0$, $uv^iw \in L(G)$.

By commutativity we can place the v's after w, and we get:

$\ldots \exists\, u, v, w \in \Sigma^*$, *such that*
$z = uwv$,
$v \neq \varepsilon$,
$|wv| \leq n$, *and*
$\forall\, i \geq 0$, $uwv^i \in L(G)$.

Let p denote $|uw|$ and q denote $|v|$. By taking the lengths of the words, which are non-negative integers, we get:

$\ldots \exists\, p \geq 0, \exists q \geq 0, \exists w \in \Sigma^*$, *such that*
(1.1) $|z| = p + q$,
(2.1) $q > 0$,
(3*) $|w| + q \leq n$, *and*
(4.1) $\forall\, s \in \Sigma^*, \forall\, i \geq 0$, if $|s| = p + i\, q$ then $s \in L(G)$.

By Condition (2.1) we can write $\exists q$, instead of $\exists q \geq 0$. Let m denote $|w|$. Since $p = |uw|$, we have that $m \leq p$, and since $q > 0$, we can write $0 < m + q \leq n$, instead of $|w| + q \leq n$.

We get:

$\ldots \exists\, p \geq 0, \exists q$, *such that*
(1.1) $|z| = p + q$,
(2.1) $q > 0$,
(3.1) $\exists m$, *with* $0 \leq m \leq p$, *such that* $0 < m + q \leq n$, *and*
(4.1) $\forall\, s \in \Sigma^*, \forall\, i \geq 0$, if $|s| = p + i\, q$ then $s \in L(G)$. □

By Condition (3.1) of Lemma 2, we can replace Condition (2.1) of Lemma 2 by the stronger condition: $0 < q \leq n$.

Let n denote the number whose existence is asserted by Lemma 2. Let us consider the following two languages subsets of $L(G)$:

(i) $L_{<n} = \{w \in L(G) \mid |w| < n\}$ and
(ii) $L_{\geq n} = \{w \in L(G) \mid |w| \geq n\}$.

Obviously, we have that $L(G) = L_{<n} \cup L_{\geq n}$. Since $L_{<n}$ is finite, $L_{<n}$ is a regular language.

Thus, in order to show that $L(G)$ is a regular language it is enough to show, as we now do, that also $L_{\geq n}$ is a regular language.

Given any word $z \in L_{\geq n}$, we have that by Lemma 2, there exist $p_0 \geq 0$ and $q_0 > 0$ such that $z = a^{p_0 + q_0}$ and $a^{p_0} \in L(G)$ (take $i = 0$).

Since $q_0 > 0$ we have that $p_0 < |z|$. Now, if $p_0 \geq n$, starting from a^{p_0}, instead of z, we get that there exist $p_1 \geq 0$ and $q_1 > 0$ such that $a^{p_0} = a^{p_1 + q_1}$, and thus,

$$z = a^{(p_1 + q_1) + q_0}.$$

In general, there exist $p_0, q_0, p_1, q_1, p_2, q_2, \ldots, p_h, q_h$, and $h \geq 0$, such that:

$$\begin{aligned} z &= a^{p_0 + q_0} = \\ &= a^{(p_1 + q_1) + q_0} = \\ &= a^{(p_2 + q_2) + q_1 + q_0} = \\ &= \ldots = \\ &= a^{(p_h + q_h) + q_{h-1} + \ldots + q_2 + q_1 + q_0} \qquad (\dagger) \end{aligned}$$

where: $(C1)$ $p_h < n$, and $(C2)$ for all i, with $0 \leq i < h$, we have that $p_i \geq n$. (Note that, when writing the expression (†), we do *not* assume that all the q_i's are distinct.)

Since for all i, with $0 \leq i \leq h$, we have that $q_i > 0$, it is the case that for any $z \in L_{\geq n}$, we can always construct an expression of the form (†) satisfying $(C1)$ and $(C2)$.

Thus, by writing $i\,q$ instead of $\overbrace{q+\ldots+q}^{i \text{ times}}$, we have that every word $z \in L_{\geq n}$ is of the form:

$$a^{p_h \,+\, i_0 q_0 \,+\, \ldots \,+\, i_k q_k}$$

for some $k, p_h, i_0, \ldots, i_k, q_0, \ldots, q_k$ such that:

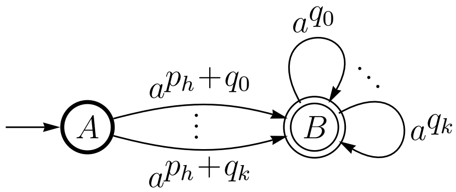
($\ell\,0$) $0 \leq k$,
($\ell\,1$) $0 \leq p_h < n$,
($\ell\,2$) $i_0 > 0, \ldots, i_k > 0$,
($\ell\,3$) $0 < q_0 \leq n$, ..., $0 < q_k \leq n$, and
($\ell\,4$) the values of $q_0, \ldots, q_k$ are *all distinct* integers and since there are at most n distinct integers r such that $0 < r \leq n$, we have that $k < n$.

Thus, the language $L_{\geq n}$, is the union of languages of the form:

$$L_{\langle p_h,\, q_0, \ldots,\, q_k\rangle} = \{a^{p_h \,+\, i_0 q_0 \,+\, \ldots \,+\, i_k q_k} \mid \; 0 \leq k \leq n, 0 \leq p_h < n, i_0 > 0, \ldots, i_k > 0, \\ 0 < q_0 \leq n, \ldots, 0 < q_k \leq n\} \;\cap\; (\{a\}^* - L_{<n})$$

Note that $L_{\geq n}$ is a *finite* union of such languages, because there exists only a finite number of tuples of the form $\langle p_h, q_0, \ldots, q_k\rangle$ such that ($\ell\,0$), ($\ell\,1$), ($\ell\,3$), and ($\ell\,4$) hold.

Note also that for any tuple of the form $\langle p_h, q_0, \ldots, q_k\rangle$ such that ($\ell\,0$), ($\ell\,1$), ($\ell\,3$), and ($\ell\,4$) hold, we have that $L_{\langle p_h,\, q_0, \ldots,\, q_k\rangle}$ is a regular language. Indeed, the finite automaton which recognizes $L_{\langle p_h,\, q_0, \ldots,\, q_k\rangle}$ is as follows:



By recalling that the class of regular languages is closed under finite union, finite intersection, and complementation, we get that $L_{\geq n}$ is a regular language.

This concludes the proof that every context-free grammar G over a terminal alphabet of cardinality 1 generates a regular language.

Note that the proof we have given, *does not* require Parikh's Lemma.
