Theorem 3.4.1. Let $G = (N, \Sigma, S, R)$ be an RTG and p be a subproper and subconsistent probability assignment for G. The following problem is NP-hard:

$$\arg\max_{\xi \in \mathrm{T}_\Sigma} \mathrm{P}(\xi \mid G, p) \ . \tag{3.3}$$

Proof sketch. Following Sima'an (2002) we reduce 3-SAT to the following decision problem: "Given a subprobabilistic RTG (G, p) and $Q \in [0, 1]$, is there $\xi \in \mathrm{T}_\Sigma$ such that $\mathrm{P}(\xi \mid G, p) > Q$?"

Let $X = \{x_1, \ldots, x_n\}$ for some $n \in \mathbb{N}$ be a set of variables and let $\mathrm{Lit}(X) = X \cup \{\neg x \mid x \in X\}$ be the set of literals. Let $k \in \mathbb{N}$ and $F = \bigwedge_{i=1}^{k} C_i$ be a 3-SAT formula where each $C_i = (u_1^i \vee u_2^i \vee u_3^i)$ for $u_j^i \in \mathrm{Lit}(X)$. For each $j \in [n]$, let n_j denote the number of occurrences of x_j or $\neg x_j$ in F. W.l.o.g. we assume that $n_j > 0$ for each $j \in [n]$.

We construct a subprobabilistic RTG (G, p) with $G = (N, \Sigma, R, S)$ as follows:

- $N = \{S\} \cup \{C_i \mid i \in [k]\} \cup \{U_j^i \mid i \in [k], j \in [3]\}$.
- $\Sigma = \{f\} \cup \{c_i \mid i \in [k]\} \cup \{u_j^i \mid i \in [k], j \in [3]\} \cup \{\top, \bot\}$ with $\mathrm{rk}(f) = k$, $\mathrm{rk}(c_i) = 3$, $\mathrm{rk}(u_j^i) = 1$, and $\mathrm{rk}(\top) = 0 = \mathrm{rk}(\bot)$.
- R is such that
  1. it contains, for each $j \in [n]$ and $b \in \{\top, \bot\}$, the rule
     $$S \to f(c_1(\xi_1^1, \ldots, \xi_3^1), \ldots, c_k(\xi_1^k, \ldots, \xi_3^k))$$
     with probability $p_j = \theta \cdot \left(\frac{1}{2}\right)^{n_j}$ such that, for each $i' \in [k]$ and $j' \in [3]$, we have
     $$\xi_{j'}^{i'} = \begin{cases} u_{j'}^{i'}(\top) & \text{if } u_{j'}^{i'} = x_j \wedge b = \top \\ u_{j'}^{i'}(\bot) & \text{if } u_{j'}^{i'} = x_j \wedge b = \bot \\ u_{j'}^{i'}(\bot) & \text{if } u_{j'}^{i'} = \neg x_j \wedge b = \top \\ u_{j'}^{i'}(\top) & \text{if } u_{j'}^{i'} = \neg x_j \wedge b = \bot \\ U_{j'}^{i'} & \text{otherwise} \end{cases} .$$
  2. it contains the rule $S \to f(C_1, \ldots, C_k)$ with probability $p_0 = 1 - 2\sum_{j \in [n]} \theta \cdot \left(\frac{1}{2}\right)^{n_j}$.
  3. for each $i \in [k]$, there are rules $C_i \to c_i(u_1^i(\top), U_2^i, U_3^i)$, $C_i \to c_i(U_1^i, u_2^i(\top), U_3^i)$, and $C_i \to c_i(U_1^i, U_2^i, u_3^i(\top))$ with probability $\frac{1}{3}$ each.
  4. for each $i \in [k]$ and $j \in [n]$, there are rules $U_j^i \to u_j^i(\top)$ and $U_j^i \to u_j^i(\bot)$ with probability $\frac{1}{2}$ each.

Note that each tree $\xi \in L(G)$ represents F with a truth value $\top$ or $\bot$ assigned to each occurrence of a literal. Each derivation d of a tree ξ is has one of following two forms:

(a) d starts with a rule of type 1 for some j followed by $3k - n_j$ rules of type 4. In this case all the assignments for literals based on x_j are consistent. The probability of this derivation is $p_j \cdot \left(\frac{1}{2}\right)^{3k-n_j} = \theta \cdot \left(\frac{1}{2}\right)^{3k}$. There are at most n derivations of this kind for ξ.

(b) d starts with a rule of type 2, followed by k rules of type 3, followed by $2k$ rules of type 4. If such an derivation exists, each clause of f contains at least one literal that assigned the value $\top$. The probability of d is $p_0 \cdot \left(\frac{1}{3}\right)^k \cdot \left(\frac{1}{2}\right)^{2k}$. There are at most 3^k derivations of this kind for ξ.

We select θ and Q such that $P(\xi \mid G, p) \geq Q$ if and only if ξ represents a variable assignment that satisfies F. Note that such a ξ must have n derivations of type (a) and at least one derivation of type (b). Hence, we choose $Q = n \cdot \theta \cdot \left(\frac{1}{2}\right)^{3k} + p_0 \cdot \left(\frac{1}{3}\right)^k \cdot \left(\frac{1}{2}\right)^{2k}$. Moreover, we set θ such that additional derivations of type (b) can not make up for missing ones of type (a):

$$
\begin{aligned}
& 3^k \cdot p_0 \cdot \left(\frac{1}{3}\right)^k \cdot \left(\frac{1}{2}\right)^{2k} < \theta \ \cdot \left(\frac{1}{2}\right)^{3k} \\
\implies \quad & (1 - 2 \sum_{j \in [n]} \theta \left(\frac{1}{2}\right)^{n_j}) < \theta \cdot \left(\frac{1}{2}\right)^k \\
\implies \quad & \frac{1}{\left(\frac{1}{2}\right)^k + 2 \sum_{j \in [n]} \left(\frac{1}{2}\right)^{n_j}} < \theta \qquad \text{(lower bound)}
\end{aligned}
$$

On the other hand, we have to construct a subproper RTG. Therefore we may choose θ such that $0 \leq p_j \leq 1$ for each $j \geq 0$. Thus, for $j = 0$ we have $-1 \leq p_0 - 1 \leq 0$, i.e., $1 \geq 2 \sum_{j \in [n]} \theta \left(\frac{1}{2}\right)^{n_j} \geq 0$ and thus:

$$
0 \leq \theta \ \leq \frac{1}{2 \sum_{j \in [n]} \left(\frac{1}{2}\right)^{n_j}} \qquad \text{(upper bound)}
$$

For $j \in [n]$, we obtain $0 \leq \theta \leq 2^{n_j}$. This upper bound on θ is less strict than the one previously stated:

$$
2^{n_j} \cdot 2 \cdot \sum_{j' \in [n]} \left(\frac{1}{2}\right)^{n_{j'}} = \underbrace{2^{n_j} \cdot 2 \cdot \left(\frac{1}{2}\right)^{n_j}}_{2} + \underbrace{2^{n_j} \cdot 2 \cdot \sum_{j' \in [n] \colon j' \neq j} \left(\frac{1}{2}\right)^{n_{j'}}}_{\geq 0} \geq 2 > 1
$$

and thus $2^{n_j} > \left(2 \sum_{j' \in [n]} \left(\frac{1}{2}\right)^{n_{j'}}\right)^{-1}$. A choice of θ that satisfies (lower bound) and (upper bound) is obviously feasible.

Note that the construction of (G, p) is polynomial in the size of F.

Now, if we can solve (3.3), then we can solve decision problem by checking if $\mathrm{P}(\hat{\xi} \mid G, p) > Q$ for $\hat{\xi} = \arg\max_{\xi \in \mathrm{T}_\Sigma} \mathrm{P}(\xi \mid G, p)$. Computing $\mathrm{P}(\hat{\xi} \mid G, p)$ can be done in polynomial time and space by intersecting G and ξ and computing the inside weight of the start symbol of the resulting grammar. Consequently, we can decide the satisfiability of F. Hence, computing (3.3) is NP-hard. □

Reference Khalil Sima'an (Aug. 2002). "Computational Complexity of Probabilistic Disambiguation". In: *Grammars* 5.2, pp. 125–151. ISSN: 1572-848X. DOI: 10.1023/A:1016340700671. URL: https://doi.org/10.1023/A:1016340700671