

# Mathematical Proofs for SYNGUAR

April 5, 2021

## 1 Preliminaries

**A starting point.** The number of examples provably sufficient to achieve the  $(\epsilon, \delta)$ -generalization is given by Blumer et al. [1]. We restate this result, which computes sample complexity as a function of  $(\epsilon, \delta)$  and the capacity (or size) of any given hypothesis space  $H$ .

**Theorem 1.1** (Sample Complexity for  $(\epsilon, \delta)$ -synthesis). *For all  $\epsilon \in (0, \frac{1}{2})$ ,  $\delta \in (0, \frac{1}{2})$ , and hypothesis space  $H$  which contains the target program  $t \in H$ , a synthesis algorithm  $\mathcal{A}(S)$  which outputs functions consistent with  $n$  i.i.d samples is an  $(\epsilon, \delta)$ -synthesizer, if*

$$n > \frac{1}{\epsilon}(\ln |H| + \ln \frac{1}{\delta})$$

*Proof.* Assume the true concept is  $f$  and all I/O examples are consistent with  $f$ . Now consider a single hypothesis  $h \in H$  first. Let  $L_{(\mathcal{D}, f)}(h)$  be the true error (expectation of 0-1 loss)  $L_{(\mathcal{D}, f)}(h) = \mathbb{E}_{x \sim \mathcal{D}} \mathbb{I}[f(x) \neq h(x)]$ , and let  $L_{(\mathcal{S}, f)}(h)$  be the empirical error (empirical average of 0-1 loss)  $L_{(\mathcal{S}, f)}(h) = \frac{1}{|S|} \sum_{x \in S} \mathbb{I}[f(x) \neq h(x)]$ , The following holds:

$$\Pr_{x \in \mathcal{D}}[\mathbb{I}[f(x) \neq h(x)] = 0 \mid L_{(\mathcal{D}, f)}(h) \geq \epsilon] \leq (1 - \epsilon)$$

Now consider a sample  $S$  with size  $n$ , the following holds:

$$\Pr_{S \in \mathcal{D}^n}[L_{(\mathcal{S}, f)}(h) = 0 \mid L_{(\mathcal{D}, f)}(h) \geq \epsilon] \leq (1 - \epsilon)^n$$

Then take union bound on all hypothesis in  $H$ :

$$\begin{aligned} & \Pr_{S \in \mathcal{D}^n}[\exists h \in H, L_{(\mathcal{S}, f)}(h) = 0 \mid L_{(\mathcal{D}, f)}(h) \geq \epsilon] \leq |H|(1 - \epsilon)^n < \delta \\ \Rightarrow n & > \frac{1}{\epsilon}(\ln |H| + \ln \frac{1}{\delta}) \text{ suffices} \end{aligned}$$

So with sample size larger than  $\frac{1}{\epsilon}(\ln |H| + \ln \frac{1}{\delta})$ , with probability at least  $1 - \delta$ , all  $h \in H$  that have true error larger than  $\epsilon$  have non-zero empirical loss and will be ruled out and any hypothesis in  $H$  that is still consistent with the sample has true error smaller than  $\epsilon$ . □

## 2 Analysis of SYNGUAR

SYNGUAR's design is motivated by being able to give a formal generalization guarantee and a bounded sample complexity. For this purpose, we state and prove the following properties:

**(P1: Termination)** SYNGUAR always terminates for a finite  $|H|$ .

---

**Algorithm 1** SYNGUAR Synthesis returns a program with error smaller than  $\epsilon$  with probability higher than  $1 - \delta$

---

```

1: procedure SYNGUAR( $\epsilon, \delta$ )
2:    $k = 10$  // tunable parameter
3:    $g \leftarrow \text{PICKSTOPPINGCOND}$ 
4:    $S' \leftarrow \emptyset, s \leftarrow 0$ 
5:    $\text{size}_H \leftarrow \text{COMPUTESIZE}(H)$ 
6:    $n \leftarrow g(\text{size}_H)$ 
7:   while  $s \leq n$  do
8:      $S' \leftarrow S' \cup \text{SAMPLE}(k)$ 
9:      $H_{S'} \leftarrow \text{UPDATEHYPOTHESIS}(S')$ 
10:     $\text{size}_{H_{S'}} \leftarrow \text{COMPUTESIZE}(H_{S'})$ 
11:     $s \leftarrow s + k$ 
12:     $n \leftarrow \min(n, s + g(\text{size}_{H_{S'}}))$ 
13:   end while
14:    $m_{H_{S'}} = \frac{1}{\epsilon}(\ln \text{size}_{H_{S'}} + \ln \frac{1}{\delta})$ 
15:    $T \leftarrow \text{SAMPLE}(m_{H_{S'}}(\epsilon, \delta))$ 
16:    $S \leftarrow S' \cup T$ 
17:   return  $f$  program in  $H_S$ 
18: end procedure

```

---

**(P2:  $(\epsilon, \delta)$  guarantees)** If SYNGUAR returns an  $f$  then  $f$  is  $\epsilon$ -far with probability  $< \delta$ .

**(P3: Sample complexity)** SYNGUAR's sample complexity is always within  $2\times$  of the optimal.

**Theorem 2.1** (P1). SYNGUAR *always terminates for a finite  $|H|$ .*

*Proof.* It suffices to prove that the sampling phase (lines 7 – 13) of SYNGUAR terminates in order to show that SYNGUAR terminates. In each iteration of the sampling phase, let  $S$  be the queue storing the user-provided examples,  $S_i = S_{i-1} \cup \{s_1, \dots, s_k\}$ . For each  $S_i$ ,  $H_{S_i}$  determines the set of consistent hypothesis that satisfy  $S_i$ . Let  $N_i$  be the number of I/O examples needed for generalization after iteration  $i$ . For iterations  $i$  and  $j$  where  $i < j$  and  $\forall g : \mathbb{N} \rightarrow \mathbb{Z}$  such that  $g$  is monotonically non-decreasing, the following holds:

$$\begin{aligned}
S_i \subset S_j &\Rightarrow |H_{S_j}| \leq |H_{S_i}| \Rightarrow g(|H_{S_j}|) \leq g(|H_{S_i}|) \\
&\Rightarrow N_j \leq N_i \text{ (see line 10 in Alg. 1)}
\end{aligned}$$

Therefore, if  $N_1 = g(|H|)$  then in the worst case the loop will terminate at some iteration  $p$  such that  $|S_p| \geq N_1$ .  $\square$

**Theorem 2.2** (P2). *If SYNGUAR  $(\epsilon, \delta)$  returns the synthesized program  $f$  then  $f$  is  $\epsilon$ -far with probability  $< \delta$ .*

*Proof.* By Theorem 2.1, we know that the sampling phase terminates with  $S'$  samples (see line 14). In lines 14 – 16 SYNGUAR samples an additional number of I/O examples required to generalize and then synthesizes a program after seeing the additional samples. Therefore, Theorem 2.2 follows from Theorem 1.1.  $\square$

In order to prove the last property, we define a new quantity  $\omega(Q)$ . It is the smallest sample size taken by SYNGUAR  $(\epsilon, \delta)$  for any non-decreasing  $g$  used for a sequence of I/O examples  $Q$ .

**Definition 1** (Smallest dynamic sample size). *For any infinite sampled sequence of examples  $Q$ , let  $\text{PREFIX}(Q, g)$  be the prefix of  $Q$  at which SYNGUAR  $(\epsilon, \delta)$  terminates. Then,*

$$\omega(Q) = \inf\{|m_g| : \forall g, m_g = \text{PREFIX}(Q, g)\}$$

**Theorem 2.3** (P3). SYNGUAR uses no more than  $2\omega(Q)$  examples on any  $Q$  with  $g_0(x) = \max\{0, \frac{1}{\epsilon}(\ln(x) - \ln(\frac{1}{\delta}))\}$ .

To save space, we provide the proof in supplementary material.

*Proof.* Let  $S' \subset Q$  be the samples in sampling phase. Let  $P$  be the samples when  $\omega(Q) = \text{PREFIX}(Q, g)$  for some  $g$  and let us call this the best stopping point for SYNGUAR's sampling phase on  $Q$ . Then  $\omega(Q) = |P| + \frac{1}{\epsilon}(\ln \text{size}_{H_P} + \ln \frac{1}{\delta})$  where  $\text{size}_{H_P} = \text{COMPUTESIZE}(H_P)$ .

There are two cases to analyze. 1) If SYNGUAR's phase 1 finishes before the best stopping point and 2) if SYNGUAR's phase 1 finishes after the best stopping point. We observe that in both cases the  $m_{g_0} \leq \omega(Q)$ . The full proof is the following:

Let  $g_0(x) = \max\{0, \frac{1}{\epsilon}(\ln(x) - \ln(1/\delta))\}$ ,  $\gamma(Q) = \text{PREFIX}(Q, g_0)$  and  $S' \subset Q$  be the samples in sampling phase.

If  $P$  be the samples for best stopping point then

$$\omega(Q) = |P| + \frac{1}{\epsilon}(\ln |H_P| + \ln \frac{1}{\delta})$$

Case 1: SYNGUAR using  $g_0$  is stopping earlier in phrase 1 than the best possible stopping point,  $|S'| < |P|$

$$\begin{aligned} \gamma(Q) - 2 \cdot \omega(Q) &= |S'| - 2 \cdot |P| + \frac{1}{\epsilon} \cdot (\log |H_{S'}| - 2 \cdot \log |H_P|) \\ &\quad - \frac{1}{\epsilon} \cdot \log \frac{1}{\delta} \\ &\leq -|P| - \frac{1}{\epsilon} \cdot \log \frac{1}{\delta} + \frac{1}{\epsilon} \cdot (\log |H_{S'}|) \\ &\text{since } |S'| < |P| \end{aligned}$$

Observe that, for  $g_0(x) = \max\{0, \frac{1}{\epsilon}(\ln(x) - \ln(1/\delta))\}$  the  $|S'| \geq \frac{1}{\epsilon} \cdot (\log |H_{S'}| - \log \frac{1}{\delta})$  (see, step 9 in Algorithm 2)

$$\gamma(Q) - 2 \cdot \omega(Q) \leq 0$$

Case 2: SYNGUAR using  $g_0$  is stopping after the best possible stopping point in phrase 1,  $|H_P| \geq |H_{S'}|$ .

Observe that  $|S'| \leq \omega(Q)$ . Because for SYNGUAR using  $g_0$ , after seeing  $P$ , it will take no more than  $g_0(|H_P|) = \frac{1}{\epsilon}(\ln |H_P| - \ln \frac{1}{\delta})$  examples in phase 1.

$$\begin{aligned} |S'| &\leq |P| + g_0(|H_P|) \\ \omega(Q) &= |P| + \frac{1}{\epsilon}(\ln |H_P| + \ln \frac{1}{\delta}) \\ \implies |S'| &\leq \omega(Q) \end{aligned}$$

Now,

$$\begin{aligned} \gamma(Q) - 2 \cdot \omega(Q) &= |S'| - 2 \cdot |P| + \frac{1}{\epsilon} \cdot (\log |H_{S'}| - 2 \cdot \log |H_P|) \\ &\quad - \frac{1}{\epsilon} \cdot \log \frac{1}{\delta} \\ \text{Substituting, } |S'| &\leq |P| + \frac{1}{\epsilon}(\log |H_P| + \log \frac{1}{\delta}) \\ &\leq -|P| + \frac{1}{\epsilon} \cdot (\log |H_{S'}| - \log |H_P|) \\ &\leq 0 \text{ since, } |H_P| \geq |H_{S'}| \end{aligned}$$

□

## References

- [1] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K Warmuth. Occam's razor. *Information processing letters*, 24(6):377–380, 1987.