Appendix for Self-Bounded Prediction Suffix Tree via Approximate String Matching

A. Proof of Lemma 2

Lemma 5. *Let* $\omega(i,k) = ((1-\xi)^k \lambda^i \exp(-\lambda)/i!)^{1/2}$. *Given binary sequence* $\mathbf{y}_1^{t-1}$ *and arbitrary suffix tree* $\mathcal{T}$ *equipped with Hamming distance metric,* $\sum_{i=1}^{t-1}\sum_{k=0}^{\epsilon}\sum_{\mathbf{s}:\mathbf{s}\in\mathcal{T},d(\mathbf{s},\mathbf{y}_{t-i}^{t-1})=k}\omega^2(i,k) \le -e^{-\lambda} + e^{\lambda(1-\xi)}\Gamma(t,-\lambda(-2+\xi))/\Gamma(t)$ *for all* $\lambda > 0$, $\epsilon \ge 0$, *and* $0 < \xi < 1$.

Proof. Given a binary sequence of length i, there are at most $\binom{i}{k}$ possible suffixes that are exactly k-Hamming distance away from the original sequence if $i \ge k$, otherwise 0.[2] Therefore the sum of ω can be bounded by the number of possible approximate suffixes as

$$\begin{aligned}
\sum_{i=1}^{t-1}\sum_{k=0}^{\epsilon}\sum_{\substack{\mathbf{s}:\mathbf{s}\in\mathcal{T},\\ \mathrm{d}(\mathbf{s},\mathbf{y}_{t-i}^{t-1})=k}}\omega^2(i,k) &\le \sum_{i=1}^{t-1}\sum_{k=0}^{\epsilon}\binom{i}{k}\omega^2(i,k) \\
&\le \sum_{i=1}^{t-1}\frac{e^{-\lambda}(\lambda(2-\xi))^i}{i!} \\
&= -e^{-\lambda} + \frac{e^{\lambda(1-\xi)}\Gamma(t,-\lambda(-2+\xi))}{\Gamma(t)} \qquad (16)
\end{aligned}$$

where $\Gamma(a)$ and $\Gamma(a,b)$ are Gamma function and incomplete Gamma function, respectively. □

B. Proof of Lemma 3

Lemma 6. *Let* $\omega(i,k) = ((1-\xi)^k \lambda^i \exp(-\lambda)/i!)^{1/2}$. *Given binary sequence* $\mathbf{y}_1^{t-1}$ *and arbitrary suffix tree* $\mathcal{T}$ *equipped with Hamming distance metric,* $\sum_{i=1}^{t-1}\sum_{k=0}^{\epsilon}\sum_{\mathbf{s}:\mathbf{s}\in\mathcal{T},d(\mathbf{s},\mathbf{y}_{t-i}^{t-1})=k}\omega^2(i,k) \le e^{\lambda}\Gamma(1+\epsilon,\lambda)/\Gamma(1+\epsilon)$ *for all* $\lambda > 0$, $\epsilon \ge 0$, *and* $0 < \xi < 1$.

Proof. Again, from the proof of Lemma 2, we can show

$$\begin{aligned}
\sum_{i=1}^{t-1}\sum_{k=0}^{\epsilon}\binom{i}{k}\omega^2(i,k) &= \sum_{i=1}^{t-1}\sum_{k=0}^{\epsilon}\frac{\lambda^k}{k!}\frac{\lambda^{i-k}\exp(-\lambda)}{(i-k)!}\mathbf{1}[i\ge k] \\
&= \sum_{k=0}^{\epsilon}\sum_{i=1}^{t-1}\frac{\lambda^k}{k!}\frac{\lambda^{i-k}\exp(-\lambda)}{(i-k)!}\mathbf{1}[i\ge k] \\
&\le \sum_{k=0}^{\epsilon}\frac{\lambda^k}{k!} \\
&= \frac{e^{\lambda}\Gamma(1+\epsilon,\lambda)}{\Gamma(1+\epsilon)} \qquad (17)
\end{aligned}$$

where we use $\sum_{i=0}^{\infty}\lambda^i\exp(-\lambda)/i! = 1$. □

C. Proof of Theorem 4

Proof. We use the same definition of Δ_t and $\hat{\Delta}_t$ as Theorem 1. The upper bound on $\sum_{t=1}(\Delta_t + \hat{\Delta}_t)$ in Eq. 9 and the equality on $\Delta_t + \hat{\Delta}_t$ in Eq. 13 still hold.

[2] Let $\binom{i}{k} = 0$ if $k \ge i$.

Let $\gamma = \min\{-e^{-\lambda} + e^{\lambda - \lambda\xi}, e^{\lambda}\Gamma(1+\epsilon,\lambda)/\Gamma(1+\epsilon)\}$. Given Eq. 13, Corollary 3.1 with the definition of h_t and $h^\star$ gives

$$\begin{aligned}\Delta_t + \hat{\Delta}_t \geq & -\tau_t^2\Big(||\mathbf{x}_t||^2 + \gamma\Big) \\ & - 2\tau_t y_t h_t(\mathbf{x}_t, \mathbf{y}_1^{t-1}) + 2\tau_t y_t h^\star(\mathbf{x}_t, \mathbf{y}_1^{t-1}) \\ & - 2\tau_t y_t \sum_{k=0}^{\epsilon}\sum_{i=d_t+1}^{t-1}\sum_{\substack{\mathbf{s}:\mathbf{s}\in\mathcal{T}^\star, \\ \mathrm{d}(\mathbf{s},\mathbf{y}_{t-i}^{t-1})=k}} \omega(i,k) g^\star(\mathbf{s}),\end{aligned} \tag{18}$$

where we subtract the last term after constructing $h^\star$ by its definition in Eq. 5. The magnitude of the last summations can be bounded by the Cauchy-Schwartz inequality

$$\begin{aligned}&\Big|\sum_{k=0}^{\epsilon}\sum_{i=d_t+1}^{t-1}\sum_{\substack{\mathbf{s}:\mathbf{s}\in\mathcal{T}^\star, \\ \mathrm{d}(\mathbf{s},\mathbf{y}_{t-i}^{t-1})=k}} \omega(i,k) g^\star(\mathbf{s})\Big| \\ &\leq \Big(\sum_{k=0}^{\epsilon}\sum_{i=d_t+1}^{t-1}\binom{i}{k}\omega(i,k)^2\Big)^{1/2}||g^\star|| \\ &= \Big(\sum_{k=0}^{\epsilon}\sum_{i=d_t+1}^{t-1}\frac{\lambda^k}{k!}\frac{\lambda^{i-k}\exp(-\lambda)}{(i-k)!}\Big)^{1/2}||g^\star||.\end{aligned} \tag{19}$$

Given that $d_t \geq \lceil \lambda + \epsilon \rceil$, the Chernoff bound (Hoeffding, 1963) gives an upper bound on the square root

$$\begin{aligned}\sum_{k=0}^{\epsilon}\sum_{i=d_t+1}^{t-1}\frac{\lambda^k}{k!}\frac{\lambda^{i-k}\exp(-\lambda)}{(i-k)!} &\leq \sum_{k=0}^{\epsilon}\frac{\lambda^k}{k!}e^{d_t-\lambda}\lambda^{d_t}d_t^{-d_t} \\ &= \frac{\Gamma(1+\epsilon,\lambda)}{\Gamma(1+\epsilon)}e^{d_t}\lambda^{d_t}d_t^{-d_t}\end{aligned}$$

Alternatively, we may use a tighter bound of the Poisson tail distribution (Glynn, 1987). For now, let $\bar{\Gamma} = \Gamma(1+\epsilon,\lambda)/\Gamma(1+\epsilon)$ and $u_\lambda(d_t) = \bar{\Gamma}e^{d_t}\lambda^{d_t}d_t^{-d_t}$. Plugging the upper bound into Eq. 18 and combining the lower bound of Ψ_t in Theorem 1 lead to

$$\begin{aligned}\Delta_t + \hat{\Delta}_t &\geq \Psi_t - 2\tau_t\sqrt{u_\lambda(d_t)}||g^\star|| \\ &\geq \tau_t\ell_t - \frac{1}{2}(\ell^\star)^2 - 2\tau_t\sqrt{u_\lambda(d_t)}||g^\star||\end{aligned} \tag{20}$$

Summing the lower bound over t and comparing to the upper bound in Eq. 9 yield

$$L_T \leq ||g^\star||^2 + ||w^\star||^2 + \frac{1}{2}\sum_{t=1}^{T}(\ell_t^\star)^2 + ||g^\star||P_t \tag{21}$$

where $P_t = \sum_{i=1}^{t} 2\tau_t\sqrt{u_\lambda(d_t)}$ and $L_t = \sum_{i=1}^{t}\tau_t\ell_t$.

We now use mathematical induction to prove that $P_t^2 \leq L_t$. Assume $P_{t-1}^2 \leq L_{t-1}$, and let $P_0 = L_0 = 0$. By the definition of P_t, we can expand

$$\begin{aligned}P_t^2 =& (P_{t-1} + 2\tau_t\sqrt{u_\lambda(d_t)})^2 \\ =& P_{t-1}^2 + 4\tau_t\sqrt{u_\lambda(d_t)}P_{t-1} + 4\tau_t^2 u_\lambda(d_t)\end{aligned} \tag{22}$$

If we choose minimum d_t which satisfies both $u_\lambda(d_t) \leq \big(((P_{t-1}^2 + \tau_t\ell_t)^{1/2} - P_{t-1})/2\tau_t\big)^2$ and $d_t \geq \lceil \lambda + \epsilon \rceil$, and plug this into Eq. 22, then with the inductive assumption we can show

$$P_t^2 \leq P_{t-1}^2 + \tau_t\ell_t \leq L_{t-1} + \tau_t\ell_t = L_t, \tag{23}$$

which proves the inductive argument. Note that the upper bound $u_\lambda(d_t)$ is strictly decreasing when $d_t \geq \lambda$, so we can always find the minimum d_t which satisfies both conditions when $\ell > 1/2$. Since P_t and L_t are always positive, we have $P_T \leq \sqrt{L_T}$. Combining this inequality with Eq. 21 leads to

$$\left(\sqrt{L_T}\right)^2 \leq ||g^\star||\sqrt{L_T} + ||g^\star||^2 + ||\mathbf{w}^\star||^2 + \frac{1}{2}\sum_{i=1}^{T}(\ell_t^\star)^2.$$

This equation is a quadratic inequality in $\sqrt{L_T}$. From the positive root of the quadratic equation, we get that

$$\sqrt{L_T} \leq \frac{1}{2}\Big(||g^\star|| + \Big(5||g^\star||^2 + 4||\mathbf{w}^\star||^2 + 2\sum_{t=1}^{T}(\ell_t^\star)^2\Big)^{\frac{1}{2}}\Big).$$

Since $\sqrt{a^2+b^2} \leq (a+b), (a,b \geq 0)$, the upper bound on $\sqrt{L_T}$ can be rewritten as

$$\sqrt{L_T} \leq \frac{1+\sqrt{5}}{2}||g^\star|| + ||\mathbf{w}^\star|| + \Big(\frac{1}{2}\sum_{t=1}^{T}(\ell_t^\star)^2\Big)^{\frac{1}{2}}. \tag{24}$$

If the loss ℓ_t at round t is greater than 1/2, then $\tau_t \ell_t \geq \ell_t^2/(3+\gamma)$ by Eq. 15, otherwise $\tau_t = 0$. Therefore the sum of ℓ_t^2 is less than $(3+\gamma)L_T$, which results the bound of Theorem 4. □

D. Experiments with Binary Synthetic Sequence

D.1. Sequence generated by single motif

We provide more comprehensive results on the synthetic binary data used in Subsection 5.1.

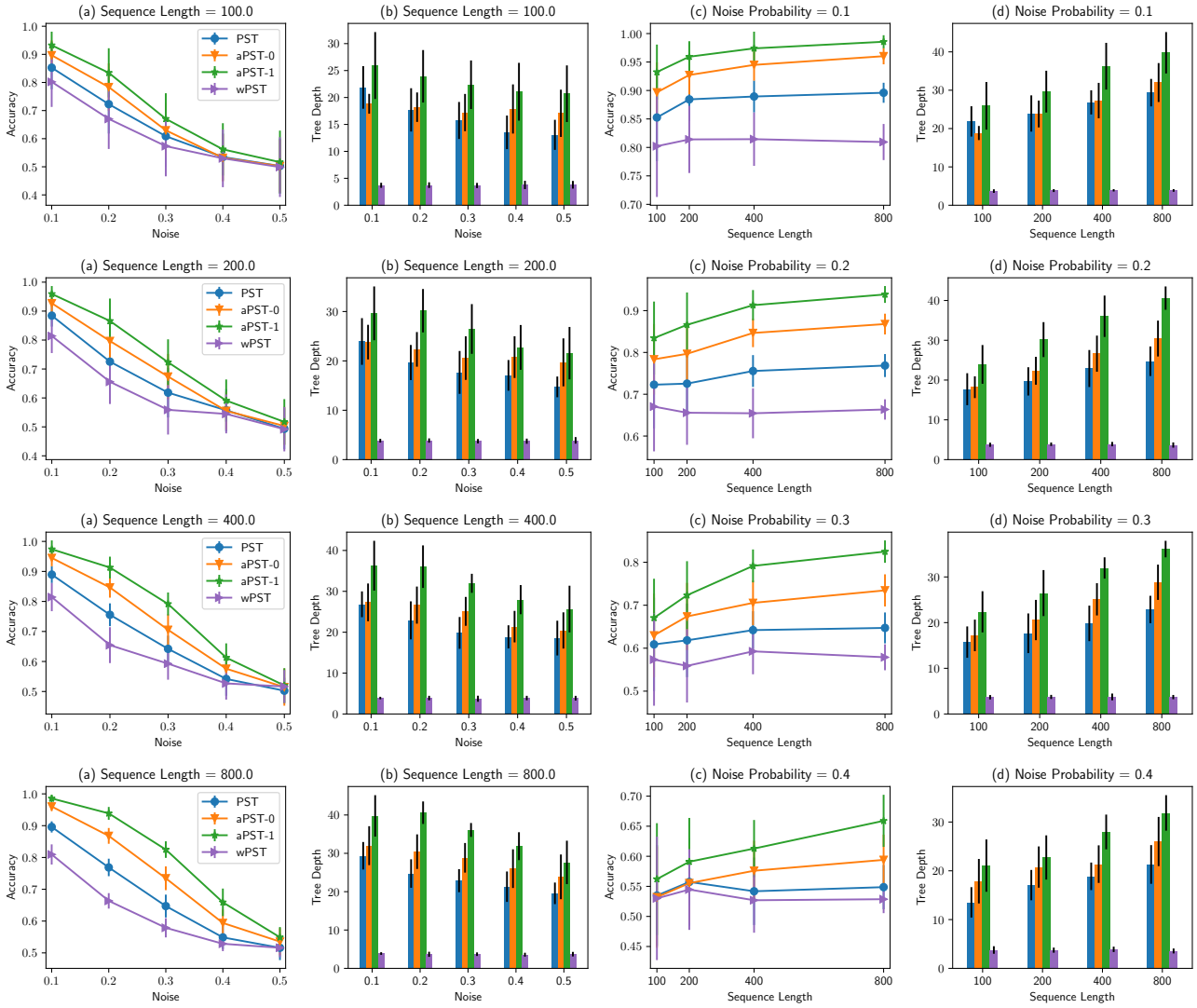

Figure 7. (a), (b): Prediction accuracy and tree depth of PSTs with respect to varying proportions of noise. Each row represents a different sequence length. (c), (d): Prediction accuracy and tree depth of PSTs with respect to varying lengths of sequence given the fixed noise level. Each row represents a different noise level.

D.2. Sequence generated by mixture of motifs

In this section, we provide experiments with more complex sequence patterns than those of the main text. For the experiments, we randomly synthesize a sequence based on two motifs: $[-1, -1, +1, +1]$ and $[+1, -1, +1, -1]$. Starting from an empty sequence, on each round, we randomly choose which motif we will append at the end of the sequence, and then add a randomly corrupted motif via a fixed noise probability. Through the above process, we generate sequences of length 100, 200, 400, 800. We randomly generate 30 sequences for each length and report the average accuracy and tree depth of each model in Figure 8.

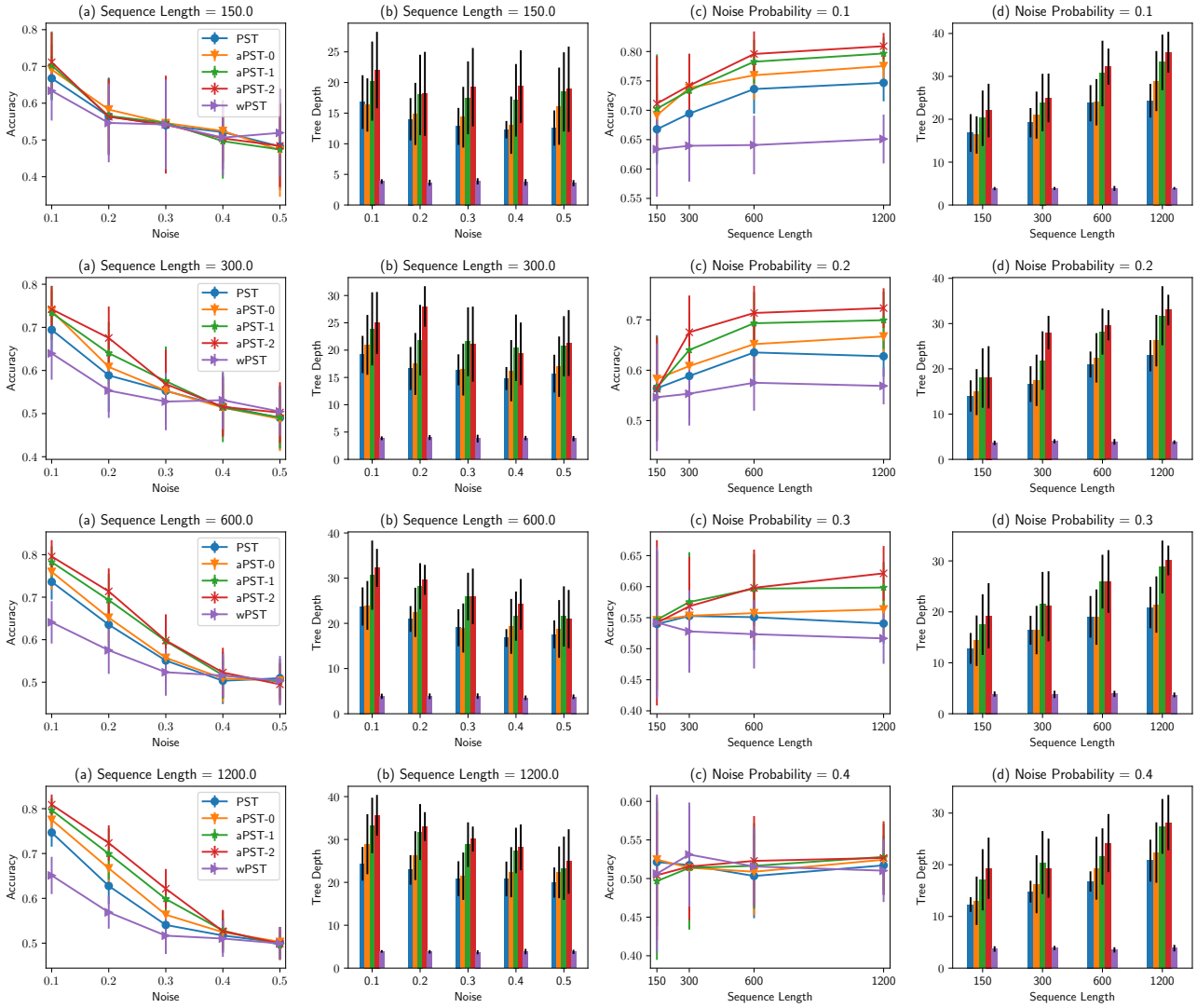

Figure 8. (a), (b): Prediction accuracy and tree depth of PSTs with respect to varying proportions of noise. Each row represents a different sequence length. (c), (d): Prediction accuracy and tree depth of PSTs with respect to varying lengths of sequence given the fixed noise level. Each row represents a different noise level.