A Proof for Proposition 1

For the notational simplicity, we omit all the index of attention head (k) and denote $W_u x$ as x. First, since both ν and p are real-valued, it suffices to consider only the real portion of e^{ix} when invoking Theorem 1. Thus, using $\text{Re}[e^{ix}] = \text{Re}[\cos(x) + i\sin(x)] = cos(x)$, we have

$$\nu(x, x') = \text{Re}[\nu(x, x')] = \int_\Omega p_\omega(\omega)\cos(\omega^\top(x - x'))d\omega.$$

Next, we have

$$\begin{aligned}
&\int_\Omega p_\omega(\omega)\cos\left(\omega^\top(x - x')\right)d\omega \\
\overset{(i)}{=}\ &\int_\Omega p_\omega(\omega)\cos\left(\omega^\top(x - x')\right)d\omega + \int_\Omega\int_0^{2\pi}\frac{1}{2\pi}p_\omega(\omega)\cos\left(\omega^\top(x + x') + 2b_u\right)db_u d\omega \\
=\ &\int_\Omega\int_0^{2\pi}\frac{1}{2\pi}p_\omega(\omega)\left[\cos\left(\omega^\top(x - x')\right) + \cos\left(\omega^\top(x + x') + 2b_u\right)\right]db_u d\omega \\
=\ &\int_\Omega\int_0^{2\pi}\frac{1}{2\pi}p_\omega(\omega)\left[2\cos(\omega^\top x + b_u)\cdot\cos(\omega^\top x' + b_u)\right]db_u d\omega \\
=\ &\int_\Omega p_\omega(\omega)\int_0^{2\pi}\frac{1}{2\pi}\left[\sqrt{2}\cos(\omega^\top x + b_u)\cdot\sqrt{2}\cos(\omega^\top x' + b_u)\right]db_u d\omega \\
=\ &\mathbb{E}\left[\phi_\omega(x)\cdot\phi_\omega(x')\right].
\end{aligned}$$

where $\phi_\omega(x) \coloneqq \sqrt{2}\cos(\omega^\top x + b_u)$, ω is sampled from p_ω, and b_u is uniformly sampled from $[0, 2\pi]$. The equation (i) holds since the second term equals to 0 as shown below:

$$\begin{aligned}
\int_\Omega\int_0^{2\pi} p_\omega(\omega)\cos\left(\omega^\top(x + x') + 2b_u\right)db_u d\omega &= \int_\Omega p_\omega(\omega)\int_0^{2\pi}\cos\left(\omega^\top(x + x') + 2b_u\right)db_u d\omega \\
&= \int_\Omega p_\omega(\omega)\cdot 0\cdot d\omega = 0.
\end{aligned}$$

Therefore, we can obtain the result in Proposition 1.

B Proof for Proposition 2

Similar to the proof in Appendix A, we omit all the index of attention head (k) and denote $W_u x$ as $x \in \mathcal{X}$ for the notational simplicity. Recall that we denote R as the radius of the Euclidean ball containing $\mathcal{X}$ in Section 3.2. In the following, we first present two useful lemmas.

Lemma 1. *Assume $\mathcal{X} \subset \mathbb{R}^d$ is compact. Let R denote the radius of the Euclidean ball containing $\mathcal{X}$, then for the kernel-induced feature mapping Φ defined in (8), the following holds for any $0 < r \le 2R$ and $\epsilon > 0$:*

$$\mathbb{P}\left\{\sup_{x, x' \in \mathcal{X}}\left|\Phi(x)^\top\Phi(x') - \nu(x, x')\right| \ge \epsilon\right\} \le 2\mathcal{N}(2R, r)\exp\left\{-\frac{D\epsilon^2}{8}\right\} + \frac{4r\sigma_p}{\epsilon}.$$

where $\sigma_p^2 = \mathbb{E}_{\omega\sim p_\omega}[\omega^\top\omega] < \infty$ is the second moment of the Fourier features, and $\mathcal{N}(R, r)$ denotes the minimal number of balls of radius r needed to cover a ball of radius R.

Proof of Lemma 1. Now, define $\Delta = \{\delta : \delta = x - x',\ , x, x' \in \mathcal{X}\}$ and note that Δ is contained in a ball of radius at most $2R$. Δ is a closed set since $\mathcal{X}$ is closed and thus Δ is a compact set. Define $B = \mathcal{N}(2R, r)$ the number of balls of radius r needed to cover Δ and let δ_j, for $j \in [B]$ denote the center of the covering balls. Thus, for any $\delta \in \Delta$ there exists a j such that $\delta = \delta_j + r'$ where $|r'| < r$.

Next, we define $S(\delta) = \Phi(x)^\top\Phi(x^\top) - \nu(x, x')$, where $\delta = x - x'$. Since S is continuously differentiable over the compact set Δ, it is L-Lipschitz with $L = \sup_{\delta\in\Delta}||\nabla S(\delta)||$. Note that if we assume $L < \frac{\epsilon}{2r}$ and for all $j \in [B]$ we have $|S(\delta_j)| < \frac{\epsilon}{2}$, then the following inequality holds for all $\delta = \delta_j + r' \in \Delta$:

$$|S(\delta)| = |S(\delta_j + r')| \le L|\delta_j - (\delta_j + r')| + |S(\delta_j)| \le rL + \frac{\epsilon}{2} < \epsilon. \tag{10}$$

The remainder of this proof bounds the probability of the events $L > \epsilon/(2r)$ and $|S(\delta_j)| \geq \epsilon/2$. Note that all following probabilities and expectations are with respect to the random variables $\omega_1, \ldots, \omega_D$.

To bound the probability of the first event, we use Proposition 1 and the linearity of expectation, which implies the key fact $\mathbb{E}[\nabla(\Phi(x)^\top \Phi(x'))] = \nabla \nu(x, x^\top)$. We proceed with the following series of inequalities:

$$\begin{aligned}
\mathbb{E}\left[L^2\right] &= \mathbb{E}\left[\sup_{\delta \in \Delta} ||\nabla S(\delta)||^2\right] \\
&= \mathbb{E}\left[\sup_{x, x' \in \mathcal{X}} ||\nabla(\Phi(x)^\top \Phi(x')) - \nabla \nu(x, x')||^2\right] \\
&\overset{(i)}{\leq} 2\mathbb{E}\left[\sup_{x, x' \in \mathcal{X}} ||\nabla(\Phi(x)^\top \Phi(x'))||^2\right] + 2 \sup_{x, x' \in \mathcal{X}} ||\nabla \nu(x, x')||^2 \\
&= 2\mathbb{E}\left[\sup_{x, x' \in \mathcal{X}} ||\nabla(\Phi(x)^\top \Phi(x'))||^2\right] + 2 \sup_{x, x' \in \mathcal{X}} ||\mathbb{E}\left[\nabla(\Phi(x)^\top \Phi(x'))\right]||^2 \\
&\overset{(ii)}{\leq} 4\mathbb{E}\left[\sup_{x, x' \in \mathcal{X}} ||\nabla(\Phi(x)^\top \Phi(x'))||^2\right],
\end{aligned}$$

where the first inequality (i) holds due to the the inequality $||a+b||^2 \leq 2||a||^2 + 2||b||^2$ (which follows from Jensen's inequality) and the subadditivity of the supremum function. The second inequality (ii) also holds by Jensen's inequality (applied twice) and again the subadditivity of supremum function. Furthermore, using a sum-difference trigonometric identity and computing the gradient with respect to $\delta = x - x'$, yield the following for any $x, x' \in \mathcal{X}$:

$$\begin{aligned}
\nabla(\Phi(x)^\top \Phi(x')) &= \nabla\left(\frac{1}{D}\sum_{i=1}^{D} \cos(\omega_i^\top (x - x'))\right) \\
&= \frac{1}{D}\sum_{i=1}^{D} \omega_i \sin(\omega_i^\top (x - x')).
\end{aligned}$$

Combining the two previous results gives

$$\begin{aligned}
\mathbb{E}[L^2] &\leq 4\mathbb{E}\left[\sup_{x, x' \in \mathcal{X}} ||\frac{1}{D}\sum_{i=1}^{D} \omega_i \sin(\omega_i^\top (x - x'))||^2\right] \\
&\leq 4 \mathop{\mathbb{E}}_{\omega_1, \ldots, \omega_D}\left[\left(\frac{1}{D}\sum_{i=1}^{D} ||\omega_i||\right)^2\right] \\
&\leq 4 \mathop{\mathbb{E}}_{\omega_1, \ldots, \omega_D}\left[\frac{1}{D}\sum_{i=1}^{D} ||\omega_i||^2\right] = 4\mathop{\mathbb{E}}_{\omega}[||\omega||^2] = 4\sigma_p^2,
\end{aligned}$$

which follows from the triangle inequality, $|\sin(\cdot)| \leq 1$, Jensen's inequality and the fact that the ω_js are drawn i.i.d. derive the final expression. Thus, we can bound the probability of the first event via Markov's inequality:

$$\mathbb{P}\left[L \geq \frac{\epsilon}{2r}\right] \leq \left(\frac{4r\sigma_p}{\epsilon}\right)^2. \tag{11}$$

To bound the probability of the second event, note that, by definition, $S(\delta)$ is a sum of D i.i.d. variables, each bounded in absolute value by $\frac{2}{D}$ (since, for all x and x', we have $|\nu(x, x')| \leq 1$ and $|\Phi(x)^\top \Phi(x')| \leq 1$), and $\mathbb{E}[S(\delta)] = 0$. Thus, by Hoeffding's inequality and the union bound, we can write

$$\mathbb{P}\left[\exists j \in [B] : |S(\delta_j)| \geq \frac{\epsilon}{2}\right] \leq \sum_{j=1}^{B} \mathbb{P}\left[|S(\delta_j)| \geq \frac{\epsilon}{2}\right] \leq 2B \exp\left(-\frac{D\epsilon^2}{8}\right). \tag{12}$$

Combining (10), (11), (12), and the definition of B we have

$$\mathbb{P}\left[\sup_{\delta \in \Delta} |S(\delta_j)| \geq \epsilon\right] \leq 2\mathcal{N}(2R, r) \exp\left\{-\frac{D\epsilon^2}{8}\right\} + \left(\frac{4r\sigma_p}{\epsilon}\right)^2.$$

□

As we can see now, a key factor in the bound of the proposition is the covering number $N(2R, r)$, which strongly depends on the dimension of the space N. In the following proof, we make this dependency explicit for one especially simple case, although similar arguments hold for more general scenarios as well.

Lemma 2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact and let R denote the radius of the smallest enclosing ball. Then, the following inequality holds:*

$$\mathcal{N}(R, r) \leq \left(\frac{3R}{r}\right)^d.$$

Proof of Lemma 2. By using the volume of balls in $\mathbb{R}^d$, we already see that $R^d/(r/3)^d = (3R/r)^d$ is a trivial upper bound on the number of balls of radius $r/3$ that can be packed into a ball of radius R without intersecting. Now, consider a maximal packing of at most $(3R/r)^d$ balls of radius $r/3$ into the ball of radius R. Every point in the ball of radius R is at distance at most r from the center of at least one of the packing balls. If this were not true, we would be able to fit another ball into the packing, thereby contradicting the assumption that it is a maximal packing. Thus, if we grow the radius of the at most $(3R/r)^d$ balls to r, they will then provide a (not necessarily minimal) cover of the ball of radius R. □

Finally, by combining the two previous lemmas, we can present an explicit finite sample approximation bound. We use lemma 1 in conjunction with lemma 2 with the following choice of r:

$$r = \left[\frac{2(6R)^d \exp(-\frac{D\epsilon^2}{8})}{\left(\frac{4\sigma_p}{\epsilon}\right)^2}\right]^{\frac{2}{d+2}},$$

which results in the following expression

$$\mathbb{P}\left[\sup_{\delta \in \Delta} |S(\delta)| \geq \epsilon\right] \leq 4\left(\frac{24R\sigma_p}{\epsilon}\right)^{\frac{2d}{d+2}} \exp\left(-\frac{D\epsilon^2}{4(d+2)}\right).$$

Since $32R\sigma_p/\epsilon \geq 1$, the exponent $2d/(d+2)$ can be replaced by 2, which completes the proof.

C Algorithm

Algorithm 1: Learning for DAPP

```
Input: The data set $X = \{\boldsymbol{x}_j\}_{j=1,\dots,n}$ with $n$ samples, where each sample $\boldsymbol{x} = \{x_i\}_{i=1}^{N_T}$ is a series of events, $N_T$ is the number of events in the time horizon $T$;
Define the number of iterations $\eta$, the number of samples in a mini-batch $M$, and the number of random Fourier features $D$;
Initialize model parameters $\boldsymbol{\theta}_0 = \{W, b, \{\theta^{(k)}, W_u^{(k)}, W_v^{(k)}\}_{k=1,\dots,K}\}$; $l = 0$;
while $l < \eta$ do
    Randomly draw $M$ sequences from $X$ denoted as $\widehat{X}_l = \{\boldsymbol{x}_j : \boldsymbol{x}_j \in \mathcal{X}\}_{j=1,\dots,M}$;
    Generate $D$ Fourier features from $p_\omega$ denoted as $\widehat{\Omega}_l = \{\omega_k \coloneqq G(z;\theta), z \sim p_z\}_{k=1,\dots,D}$;
    $\boldsymbol{\theta}_l \leftarrow$ Update $\boldsymbol{\theta}_l$ by maximizing (1) using stochastic gradient descent given $\widehat{X}_l, \widehat{\Omega}_l$;
    $l \leftarrow l + 1$;
end
```

Algorithm 2: Efficient thinning algorithm for DAPP

```
input $\boldsymbol{\theta}, T, \mathcal{M}$;
output A set of events $\mathcal{H}_t$ ordered by time.;
Initialize $\mathcal{H}_t = \emptyset$, $t = 0$, $m \sim \texttt{uniform}(\mathcal{M})$;
while $t < T$ do
    Sample $u \sim \texttt{uniform}(0, 1)$; $m \sim \texttt{uniform}(\mathcal{M})$; $D \sim \texttt{uniform}(0, 1)$;
    $x' \leftarrow (t, m')$; $\bar{\lambda} \leftarrow \lambda(x' | \boldsymbol{h}(x'))$ given history $\mathcal{H}_t$;
    $t \leftarrow t - \ln u / \bar{\lambda}$;
    $x \leftarrow (t, m)$; $\widetilde{\lambda} \leftarrow \lambda(x | \boldsymbol{h}(x))$ given history $\mathcal{H}_t$;
    if $D\bar{\lambda} > \widetilde{\lambda}$ then
        $\mathcal{H}_t \leftarrow \mathcal{H}_t \cup \{(t, m)\}$; $m' \leftarrow m$;
    end
end
```

Algorithm 3: Event selection for online attention

```
Input: data $\boldsymbol{x} = \{x_i\}_{i=1}^{\infty}$, threshold $\eta$;
Initialize $\mathscr{A}_0^{(k)} = \emptyset, k = 1, \dots, K$;
for $i = 1$ to $+\infty$. do
    for $k = 1$ to $K$. do
        $\mathscr{A}_i^{(k)} \leftarrow \mathscr{A}_{i-1}^{(k)} \cup \{x_i\}$;
        Initialize $\mathscr{S}_i^{(k)} = \emptyset$, $\bar{\nu}_j^{(k)} = 0$;
        for $j = 1$ to $i - 1$ do
            $\mathscr{S}_j^{(k)} \leftarrow \mathscr{S}_j^{(k)} \cup \widetilde{\nu}^{(k)}(x_i, x_j)$;
            $\bar{\nu}_j^{(k)} \leftarrow (\sum_{s \in \mathscr{S}_j^{(k)}} s) / |\mathscr{S}_j^{(k)}|$;
        end
        if $i > \eta$ then
            $\mathscr{A}_i^{(k)} \leftarrow \mathscr{A}_{i-1}^{(k)} \setminus \underset{x_j : t_j < t_i}{\arg\min} \left\{ \bar{\nu}_j^{(k)} \right\}$;
        end
    end
end
```