A. Proof of Lemma 1

In order to prove Lemma 1, we first establish the following lemma as a step stone.

Lemma 3. *Under (a1), (a2), (20a) and (20b) with* $\mathcal{F}_n = \{\hat{f}_n | \hat{f}_n(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}), \forall \boldsymbol{\theta} \in \mathbb{R}^{2D}\}$*, let* $\hat{f}_{RF,n}(.)$ *denote the sequence of estimates generated by our MKL algorithm with a preselected kernel* κ_n*. The following bound holds true with probability 1:*

$$\sum_{t=1}^{T} \mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) - \sum_{t=1}^{T} \mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t), y_t) \leq \frac{\|\boldsymbol{\theta}^*_n\|^2}{2\eta q_n^{min}} + \frac{\eta L^2 T}{2} \tag{25}$$

where η *is the learning rate,* L *is the Lipschitz constant in (a2),* $q_n^{min} = \min_{\forall t \in \{1,\ldots,T\}} q_{n,t}$*, and* $\boldsymbol{\theta}^*_n$ *is the corresponding parameter vector supporting the best estimator* $\hat{f}^*_{t,n}(\mathbf{x}) = (\boldsymbol{\theta}^*_n)^\top \mathbf{z}_n(\mathbf{x})$*.*

Proof. Note that OMKL-GF updates the $\boldsymbol{\theta}_{n,t}$ only if the n-th kernel is in the chosen subset. Therefore, based on (12), for any fixed $\boldsymbol{\theta}$, we find

$$\begin{aligned}\|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2 &= \|\boldsymbol{\theta}_{n,t} - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\mathcal{I}(n \in \mathcal{S}_t) - \boldsymbol{\theta}\|^2 \\ &= \|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - 2\eta \nabla^\top \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\mathcal{I}(n \in \mathcal{S}_t)(\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}) \\ &\quad + \|\eta \nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\mathcal{I}(n \in \mathcal{S}_t)\|^2. \end{aligned} \tag{26}$$

Furthermore, based on the convexity of loss function under (a1), it can be written that

$$\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) \leq \nabla^\top \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)(\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}) \tag{27}$$

Combining (26) with (27), we arrive at

$$\begin{aligned}&\left(\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\right)\mathcal{I}(n \in \mathcal{S}_t) \\ &\quad \leq \frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta} + \frac{\eta}{2}\|\nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\mathcal{I}(n \in \mathcal{S}_t)\|^2. \end{aligned} \tag{28}$$

Taking the expectation of left hand side of (28) with respect to $\mathcal{I}(n \in \mathcal{S}_t)$, we obtain

$$\begin{aligned}&\mathbb{E}[\left(\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\right)\mathcal{I}(n \in \mathcal{S}_t)] \\ &= \left(\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\right) \times 1 \times q_{n,t} + \left(\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\right) \times 0 \times (1 - q_{n,t}) \\ &= q_{n,t}\left(\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\right) \end{aligned} \tag{29}$$

where $q_{n,t}$ is the probability that the n-th kernel is in the chosen subset of kernels. Moreover, for the expectation of right hand side of (28), we have

$$\begin{aligned}&\mathbb{E}\left[\frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta} + \frac{\eta}{2}\|\nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\mathcal{I}(n \in \mathcal{S}_t)\|^2\right] \\ &= \frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta} + \frac{\eta q_{n,t}}{2}\|\nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\|^2. \end{aligned} \tag{30}$$

From (28), (29) and (30), we can conclude that

$$\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) \leq \frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta q_{n,t}} + \frac{\eta}{2}\|\nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\|^2. \tag{31}$$

Summing (31) over $t = 1, \ldots, T$, we obtain

$$\sum_{t=1}^{T}\left(\mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\right) \leq \sum_{t=1}^{T}\frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta q_{n,t}} + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\|^2. \tag{32}$$

Let $q_n^{\min} = \min_{\forall t \in \{1,...,T\}} q_{n,t}$. Based on (a2), the right hand side of (32) can be bounded by

$$\sum_{t=1}^{T} \frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta q_{n,t}} + \frac{\eta}{2}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t)\|^2 \le \sum_{t=1}^{T} \frac{\|\boldsymbol{\theta}_{n,t} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,t+1} - \boldsymbol{\theta}\|^2}{2\eta q_n^{\min}} + \frac{\eta}{2}\sum_{t=1}^{T} L^2$$
$$= \frac{\|\boldsymbol{\theta}_{n,1} - \boldsymbol{\theta}\|^2 - \|\boldsymbol{\theta}_{n,T+1} - \boldsymbol{\theta}\|^2}{2\eta q_n^{\min}} + \frac{\eta L^2 T}{2} \quad (33)$$

where L is the Lipschitz constant. Using the facts that $\boldsymbol{\theta}_{n,1} = \mathbf{0}$ and non-negativity of $\|\boldsymbol{\theta}_{n,T+1} - \boldsymbol{\theta}\|^2$, from (32) and (33) we can conclude that

$$\sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) - \sum_{t=1}^{T} \mathcal{L}(\boldsymbol{\theta}^\top \mathbf{z}_n(\mathbf{x}_t), y_t) \le \frac{\|\boldsymbol{\theta}\|^2}{2\eta q_n^{\min}} + \frac{\eta L^2 T}{2}. \quad (34)$$

By choosing $\boldsymbol{\theta} = \boldsymbol{\theta}_n^*$ such that $\hat{f}_{t,n}^*(\mathbf{x}) = (\boldsymbol{\theta}_n^*)^\top \mathbf{z}_n(\mathbf{x})$, we arrive at

$$\sum_{t=1}^{T} \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t) - \sum_{t=1}^{T} \mathcal{L}(\hat{f}_{t,n}^*(\mathbf{x}_t), y_t) \le \frac{\|\boldsymbol{\theta}_n^*\|^2}{2\eta q_n^{\min}} + \frac{\eta L^2 T}{2} \quad (35)$$

where $\hat{f}_{\text{RF},n}(\mathbf{x}_t) = \boldsymbol{\theta}_{n,t}^\top \mathbf{z}_n(\mathbf{x}_t)$. □

Lemma 4. *Under (a1) and (a2), the following holds*

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t), y_t)] - \sum_{t=1}^{T} \mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) \le \frac{2^b}{\eta} \ln N + \eta_e JT + \frac{\eta}{2^{b+1}} \sum_{t=1}^{T}\sum_{n=1}^{N} \frac{1}{q_{n,t}} \quad (36)$$

where η is the learning rate, η_e is the exploration rate, $b = \lfloor \log_2(J) \rfloor$, $q_{n,t} = \sum_{j=1}^{J} p_{j,t}\left(1 - (1 - p_{t,j}^{(\kappa_n)})^M\right)$ and N denotes the number of kernels.

Proof. Let $W_t = \sum_{n=1}^{N} w_{n,t}$. For any t we find

$$\frac{W_{t+1}}{W_t} = \sum_{j=1}^{J} p_{j,t} \frac{W_{t+1}}{W_t} = \sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{w_{n,t+1}}{W_t} = \sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{w_{n,t}}{W_t} \exp(-\frac{\eta}{2^b}\ell_{n,t}). \quad (37)$$

Based on (17), we have

$$\frac{w_{n,t}}{W_t} = \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1 - \eta_e^j}, \forall j \in \{1, ..., J\}. \quad (38)$$

Combining (37) with (38) obtains

$$\frac{W_{t+1}}{W_t} = \sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1 - \eta_e^j} \exp(-\frac{\eta}{2^b}\ell_{n,t}). \quad (39)$$

Using the inequality $e^{-x} \le 1 - x + \frac{1}{2}x^2, \forall x \ge 0$, (39) leads to

$$\frac{W_{t+1}}{W_t} \le \sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1 - \eta_e^j} \left(1 - \frac{\eta}{2^b}\ell_{n,t} + \frac{1}{2}(\frac{\eta}{2^b}\ell_{n,t})^2\right). \quad (40)$$

Taking logarithm from both sides of inequality (40), and use the fact that $1 + x \le e^x$, we have

$$\ln \frac{W_{t+1}}{W_t} \le \sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1 - \eta_e^j} \left(-\frac{\eta}{2^b}\ell_{n,t} + \frac{1}{2}(\frac{\eta}{2^b}\ell_{n,t})^2\right). \quad (41)$$

Summing (41) over t from 1 to T results in

$$\ln \frac{W_{T+1}}{W_1} \leq \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1-\eta_e^j} \left(-\frac{\eta}{2^b}\ell_{n,t} + \frac{1}{2}(\frac{\eta}{2^b}\ell_{n,t})^2\right). \tag{42}$$

Furthermore, recall the updating rule of $w_{n,T+1}$ in (13), for any n we have

$$\ln \frac{W_{T+1}}{W_1} \geq \ln \frac{w_{n,T+1}}{W_1} = -\ln N - \sum_{t=1}^{T} \frac{\eta}{2^b}\ell_{n,t}. \tag{43}$$

Combining (42) with (43) results in

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)}}{1-\eta_e^j}(\frac{\eta}{2^b}\ell_{n,t}) - \sum_{t=1}^{T}\frac{\eta}{2^b}\ell_{n,t}\\ \leq& \ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\frac{\eta_e^j}{N}}{1-\eta_e^j}(\frac{\eta}{2^b}\ell_{n,t}) + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1-\eta_e^j}\left(\frac{1}{2}(\frac{\eta}{2^b}\ell_{n,t})^2\right).\end{aligned} \tag{44}$$

Multiplying both sides by $(1-\eta_e^J)\frac{2^b}{\eta}$, we arrive at

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} p_{t,j}^{(\kappa_n)} \frac{1-\eta_e^J}{1-\eta_e^j}\ell_{n,t} - \sum_{t=1}^{T}(1-\eta_e^J)\ell_{n,t}\\ \leq&(1-\eta_e^J)\frac{2^b}{\eta}\ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e^j(1-\eta_e^J)}{N(1-\eta_e^j)}\ell_{n,t} + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{(1-\eta_e^J)(p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N})}{1-\eta_e^j}(\frac{\eta}{2^{b+1}}\ell_{n,t}^2).\end{aligned} \tag{45}$$

Also, using the fact that $0 < \eta_e \leq 1$ we can conclude that $1-\eta_e^J < 1$ and for all $j \geq 1$, $\eta_e^j \leq \eta_e$, the RHS of (45) can be upper bounded by

$$\begin{aligned}&(1-\eta_e^J)\frac{2^b}{\eta}\ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e^j(1-\eta_e^J)}{N(1-\eta_e^j)}\ell_{n,t} + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{(1-\eta_e^J)(p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N})}{1-\eta_e^j}(\frac{\eta}{2^{b+1}}\ell_{n,t}^2)\\ \leq& \frac{2^b}{\eta}\ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e(1-\eta_e^J)}{N(1-\eta_e)}\ell_{n,t} + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1-\eta_e^j}(\frac{\eta}{2^{b+1}}\ell_{n,t}^2).\end{aligned} \tag{46}$$

Since $1-\eta_e^J = (1-\eta_e)(1+\ldots+\eta_e^{J-1})$ and $\eta_e \leq 1$, the following bound holds for the second term on the RHS of (46)

$$\begin{aligned}\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e(1-\eta_e^J)}{N(1-\eta_e)}\ell_{n,t} =& \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e(1+\ldots+\eta_e^{J-1})}{N}\ell_{n,t}\\ \leq& \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e J}{N}\ell_{n,t}.\end{aligned} \tag{47}$$

Meanwhile, as $\eta_e^J \leq \eta_e^j$ for all j, $1 \leq j \leq J$, the LHS of (45) can be bounded by

$$\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} p_{t,j}^{(\kappa_n)} \frac{1-\eta_e^J}{1-\eta_e^j}\ell_{n,t} - \sum_{t=1}^{T}(1-\eta_e^J)\ell_{n,t} \geq \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} p_{t,j}^{(\kappa_n)}\ell_{n,t} - \sum_{t=1}^{T}\ell_{n,t}. \tag{48}$$

Combining (45), (46), (47) and (48), we can conclude that

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} p_{t,j}^{(\kappa_n)}\ell_{n,t} - \sum_{t=1}^{T}\ell_{n,t}\\ \leq&\frac{2^b}{\eta}\ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{\eta_e J}{N}\ell_{n,t} + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} \frac{p_{t,j}^{(\kappa_n)} - \frac{\eta_e^j}{N}}{1-\eta_e^j}(\frac{\eta}{2^{b+1}}\ell_{n,t}^2).\end{aligned} \tag{49}$$

Recall the probability of observing the loss of n-th kernel at time t given in (18), the expected first and second moments of $\ell_{n,t}$ in (14) given the losses incurred up to time instant $t-1$, i.e., $\{\mathcal{L}(\hat{f}_\tau(\mathbf{x}_\tau), y_\tau)\}_{\tau=1}^{t-1}$ can be written as

$$\mathbb{E}[\ell_{n,t}] = \sum_{j=1}^{J} p_{j,t}\left(1-(1-p_{t,j}^{(\kappa_n)})^M\right)\frac{\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t)}{q_{n,t}} = \mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) \tag{50a}$$

$$\mathbb{E}[\ell_{n,t}^2] = \sum_{j=1}^{J} p_{j,t}\left(1-(1-p_{t,j}^{(\kappa_n)})^M\right)\frac{\mathcal{L}^2(\hat{f}_{RF,n}(\mathbf{x}_t), y_t)}{q_{n,t}^2} = \frac{\mathcal{L}^2(\hat{f}_{RF,n}(\mathbf{x}_t), y_t)}{q_{n,t}} \le \frac{1}{q_{n,t}}. \tag{50b}$$

Based on (50b), the third term in the right hand side of (49) can be bounded as follows

$$\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{p_{t,j}^{(\kappa_n)}-\frac{\eta_e^j}{N}}{1-\eta_e^j}\left(\frac{\eta}{2^{b+1}}\ell_{n,t}^2\right) \le \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{p_{t,j}^{(\kappa_n)}-\frac{\eta_e^j}{N}}{1-\eta_e^j}\left(\frac{\eta}{2^{b+1}q_{n,t}}\right). \tag{51}$$

Taking the expected value of (49) at each time t given$\{\mathcal{L}(\hat{f}_\tau(\mathbf{x}_\tau), y_\tau)\}_{\tau=1}^{t-1}$ and combining with (50a) and (51) we have

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N} p_{t,j}^{(\kappa_n)}\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) - \sum_{t=1}^{T}\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t)\\ \le&\frac{2^b}{\eta}\ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{\eta_e J}{N}\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{p_{t,j}^{(\kappa_n)}-\frac{\eta_e^j}{N}}{1-\eta_e^j}\left(\frac{\eta}{2^{b+1}q_{n,t}}\right).\end{aligned} \tag{52}$$

Since $\frac{w_{n,t}}{W_t} = \frac{p_{t,j}^{(\kappa_n)}-\frac{\eta_e^j}{N}}{1-\eta_e^j} \le 1$, replace $\frac{p_{t,j}^{(\kappa_n)}-\frac{\eta_e^j}{N}}{1-\eta_e^j} \le 1$ by 1, the inequality in (52) still holds

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N} p_{t,j}^{(\kappa_n)}\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) - \sum_{t=1}^{T}\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t)\\ \le&\frac{2^b}{\eta}\ln N + \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{\eta_e J}{N}\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t), y_t) + \frac{\eta}{2^{b+1}}\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{1}{q_{n,t}}.\end{aligned} \tag{53}$$

Also, using the fact that $\sum_{j=1}^{L} p_{j,t} = 1$, for the third term in the right hand side of (53) we have

$$\frac{\eta}{2^{b+1}}\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{1}{q_{n,t}} = \frac{\eta}{2^{b+1}}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{1}{q_{n,t}}. \tag{54}$$

Furthermore, based on that $\mathcal{L}(\hat{f}_{\mathrm{RF},n}(\mathbf{x}_t), y_t) \le 1$ in (a2), the following inequality holds

$$\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{\eta_e J}{N}\mathcal{L}(\hat{f}_{\mathrm{RF},n}(\mathbf{x}_t), y_t) \le \sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N}\frac{\eta_e J}{N} = \eta_e JT. \tag{55}$$

From (53), (54) and (55), we can conclude that

$$\begin{aligned}&\sum_{t=1}^{T}\sum_{j=1}^{J} p_{j,t}\sum_{n=1}^{N} p_{t,j}^{(\kappa_n)}\mathcal{L}(\hat{f}_{\mathrm{RF},n}(\mathbf{x}_t), y_t) - \sum_{t=1}^{T}\mathcal{L}(\hat{f}_{\mathrm{RF},n}(\mathbf{x}_t), y_t)\\ \le&\frac{2^b}{\eta}\ln N + \eta_e JT + \frac{\eta}{2^{b+1}}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{1}{q_{n,t}}.\end{aligned} \tag{56}$$

According to the procedure of generating the graph G_t which is presented in Algorithm 1, for each selective node $v_j^{(c)}$ a subset of kernels is chosen using PMF $p_{t,j}^{(\kappa)}$ in M independent trials. In fact, a subset of kernels is assigned to each node

$v_j^{(c)}$ in M independent trials and in each trial one kernel is assigned and its associated entry in the sub-adjacency matrix A becomes 1. Now, let b_n represents the frequency that n-th kernel is chosen in M independent trials. Thus, $\{b_n\}_{n=1}^N$ can be viewed as the solution to the following linear equation

$$b_1 + \ldots + b_N = M, \text{ s.t. } b_n \geq 0, b_n \in \mathbb{N} \tag{57}$$

where $\mathbb{N}$ denotes the set of natural numbers. There are $\binom{N+M-1}{N}$ different solutions for (57). Let, $\{b_{n,k}\}_{n=1}^N$ denotes k-th set of solution for (57). Based on Jensen's inequality, for the expected value of $\mathcal{L}(\hat{f}_t(\mathbf{x}_t), y_t)$ we have

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t), y_t)] &= \sum_{j=1}^{J} p_{j,t} \sum_{k=1}^{\binom{N+M-1}{N}} \left(\prod_{n=1}^{N} (p_{t,j}^{(\kappa_n)})^{b_{n,k}} \right) \mathcal{L}(\sum_{n\in\mathcal{S}_t} \bar{w}_{n,t} \hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t) \\ &\leq \sum_{j=1}^{J} p_{j,t} \sum_{k=1}^{\binom{N+M-1}{N}} \left(\prod_{n=1}^{N} (p_{t,j}^{(\kappa_n)})^{b_{n,k}} \right) \sum_{n\in\mathcal{S}_t} \bar{w}_{n,t} \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t).\end{aligned} \tag{58}$$

Also, considering (58) and the fact that $\bar{w}_{n,t} \leq 1$, we can conclude that

$$\begin{aligned}\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t), y_t)] &\leq \sum_{j=1}^{J} p_{j,t} \sum_{k=1}^{\binom{N+M-1}{N}} \left(\prod_{n=1}^{N} (p_{t,j}^{(\kappa_n)})^{b_{n,k}} \right) \sum_{n\in\mathcal{S}_t} \bar{w}_{n,t} \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t) \\ &\leq \sum_{j=1}^{J} p_{j,t} \sum_{k=1}^{\binom{N+M-1}{N}} \left(\prod_{n=1}^{N} (p_{t,j}^{(\kappa_n)})^{b_{n,k}} \right) \sum_{n\in\mathcal{S}_t} \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t).\end{aligned} \tag{59}$$

Note that the number of ways to solve (57) when n-th kernel is chosen for at least one time equals to the number of ways to solve the following problem

$$\tilde{b}_{1,n} + \ldots + \tilde{b}_{N,n} = M - 1, \text{ s.t. } \tilde{b}_{m,n} \geq 0, \ \tilde{b}_{m,n} \in \mathbb{N}. \tag{60}$$

There are $\binom{N+M-2}{N}$ different solutions for (60). Let $\{\tilde{b}_{m,n}^{(k)}\}_{n=1}^N$ denotes k-th set of solution for (60). Therefore, based on this, from (59) we can conclude the following equality

$$\begin{aligned}&\sum_{j=1}^{J} p_{j,t} \sum_{k=1}^{\binom{N+M-1}{N}} \left(\prod_{n=1}^{N} (p_{t,j}^{(\kappa_n)})^{b_{n,k}} \right) \sum_{n\in\mathcal{S}_t} \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t) \\ =&\sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} p_{t,j}^{(\kappa_n)} \sum_{k=1}^{\binom{N+M-2}{N}} \left(\prod_{m=1}^{N} (p_{t,j}^{(\kappa_m)})^{\tilde{b}_{m,n}^{(k)}} \right) \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t)\end{aligned} \tag{61}$$

where $\sum_{k=1}^{\binom{N+M-2}{N}} \left(\prod_{m=1}^{N} (p_{t,j}^{(\kappa_m)})^{\tilde{b}_{m,n}^{(k)}} \right)$ is the total probability of all $\binom{N+M-2}{N}$ possible solutions of (60). Therefore, $\sum_{k=1}^{\binom{N+M-2}{N}} \left(\prod_{m=1}^{N} (p_{t,j}^{(\kappa_m)})^{\tilde{b}_{m,n}^{(k)}} \right) = 1$. Substituting (61) into (58), we obtain

$$\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t), y_t)] \leq \sum_{j=1}^{J} p_{j,t} \sum_{n=1}^{N} p_{t,j}^{(\kappa_n)} \mathcal{L}(\hat{f}_{\text{RF},n}(\mathbf{x}_t), y_t). \tag{62}$$

Combining (56) with (62) leads to

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t), y_t)] - \sum_{t=1}^{T} \mathcal{L}(\hat{f}_{n,t}(\mathbf{x}_t), y_t) \leq \frac{2^b}{\eta} \ln N + \eta_e JT + \frac{\eta}{2^{b+1}} \sum_{t=1}^{T} \sum_{n=1}^{N} \frac{1}{q_{n,t}} \tag{63}$$

which concludes to proof of Lemma 4. □

From (25) in Lemma 3 and (36) in Lemma 4, we conclude that

$$\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t),y_t)]-\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)\leq\frac{2^b}{\eta}\ln N+\frac{\|\boldsymbol{\theta}^*_n\|^2}{2\eta q_n^{\min}}+\frac{\eta L^2T}{2}+\eta_e JT+\frac{\eta}{2^{b+1}}\sum_{t=1}^{T}\sum_{n=1}^{N}\frac{1}{q_{n,t}}. \tag{64}$$

Furthermore, based on (18) we can write

$$q_{n,t}=\sum_{j=1}^{J}p_{j,t}\left(1-(1-p_{t,j}^{(\kappa_n)})^M\right)=\sum_{j=1}^{J}p_{j,t}p_{t,j}^{(\kappa_n)}\left(1+\ldots+(1-p_{t,j}^{(\kappa_n)})^{M-1}\right)\geq\sum_{j=1}^{J}p_{j,t}p_{t,j}^{(\kappa_n)}. \tag{65}$$

From (65) and the facts that $p_{j,t}>\frac{\eta_e}{J}$ and $p_{t,j}^{(\kappa_n)}>\frac{\eta_e^j}{N}$, the following inequality can be concluded

$$q_{n,t}\geq\sum_{j=1}^{J}p_{j,t}p_{t,j}^{(\kappa_n)}>p_{1,t}p_{t,1}^{(\kappa_n)}>\frac{\eta_e^2}{NJ}. \tag{66}$$

Therefore, we find $q_n^{\min}>\frac{\eta_e^2}{NJ}$. Combining (64) and (66) we can conclude that

$$\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t),y_t)]-\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)<\frac{2^b}{\eta}\ln N+\frac{\|\boldsymbol{\theta}^*_n\|^2NJ}{2\eta\eta_e^2}+\frac{\eta L^2T}{2}+\eta_e JT+\frac{\eta N^2JT}{2^{b+1}\eta_e^2}. \tag{67}$$

Hence, Lemma 1 is proved.

B. Proof of Theorem 2

To prove Theorem 2, the following lemma is exploited (Shen et al., 2019)

Lemma 5. *For the optimal function estimator (19) in $\mathcal{H}_n$ expressed as $f_n^*(\mathbf{x}):=\sum_{t=1}^{T}\alpha^*_{n,t}\kappa_n(\mathbf{x},\mathbf{x}_t)$ and its RF-based approximant $\hat{f}^*_{t,n}(\mathbf{x},\mathbf{x}_t)=\sum_{t=1}^{T}\alpha^*_{n,t}\mathbf{z}_n^\top(\mathbf{x})\mathbf{z}_n(\mathbf{x}_t)$, the following bound holds with probability at least $1-2^8(\frac{\sigma_n}{\epsilon})^2\exp(-\frac{D\epsilon^2}{4d+8})$*

$$\left|\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\sum_{t=1}^{T}\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\right|\leq\epsilon LTC \tag{68}$$

where the equality happens if we have $C:=\max_n\sum_{t=1}^{T}|\alpha^*_{n,t}|$.

Proof. For a given shift invariant kernel κ_n, the maximum point-wise error of the random feature kernel approximant is uniformly bounded with probability at least $1-2^8(\frac{\sigma_n}{\epsilon})^2\exp(-\frac{D\epsilon^2}{4d+8})$, by (Rahimi & Recht, 2007)

$$\sup_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}}|\mathbf{z}_n(\mathbf{x}_i)^\top\mathbf{z}_n(\mathbf{x}_j)-\kappa_n(\mathbf{x}_i,\mathbf{x}_j)|<\epsilon \tag{69}$$

Furthermore, using the triangle inequality we can conclude that

$$\left|\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\sum_{t=1}^{T}\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\right|\leq\sum_{t=1}^{T}\left|\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\right|. \tag{70}$$

Considering the Lipschitz continuity of the loss function we can obtain the following inequality

$$\sum_{t=1}^{T}\left|\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\right|\leq\sum_{t=1}^{T}L\left|\sum_{\tau=1}^{T}\alpha^*_{n,\tau}\mathbf{z}_n^\top(\mathbf{x}_\tau)\mathbf{z}_n(\mathbf{x}_t)-\sum_{\tau=1}^{T}\alpha^*_{n,\tau}\kappa_n(\mathbf{x}_\tau,\mathbf{x}_t)\right|. \tag{71}$$

Using the Cauchy-Schwartz inequality, we obtain

$$\sum_{t=1}^{T}L\left|\sum_{\tau=1}^{T}\alpha^*_{n,\tau}\mathbf{z}_n^\top(\mathbf{x}_\tau)\mathbf{z}_n(\mathbf{x}_t)-\sum_{\tau=1}^{T}\alpha^*_{n,\tau}\kappa_n(\mathbf{x}_\tau,\mathbf{x}_t)\right|\leq\sum_{t=1}^{T}L\sum_{\tau=1}^{T}|\alpha^*_{n,\tau}|\left|\mathbf{z}_n^\top(\mathbf{x}_\tau)\mathbf{z}_n(\mathbf{x}_t)-\kappa_n(\mathbf{x}_\tau,\mathbf{x}_t)\right| \tag{72}$$

Hence, from (70), (71) and (72) we can conclude the following inequality

$$\left|\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\sum_{t=1}^{T}\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\right|\leq\sum_{t=1}^{T}L\sum_{\tau=1}^{T}|\alpha^*_{n,\tau}|\left|\mathbf{z}_n^\top(\mathbf{x}_\tau)\mathbf{z}_n(\mathbf{x}_t)-\kappa_n(\mathbf{x}_\tau,\mathbf{x}_t)\right| \tag{73}$$

Combining (69) with (73) and considering the fact that $C:=\max_n\sum_{t=1}^{T}|\alpha^*_{n,t}|$, yields the following inequality which holds with probability at least $1-2^8(\frac{\sigma_n}{\epsilon})^2\exp(-\frac{D\epsilon^2}{4d+8})$,

$$\left|\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\sum_{t=1}^{T}\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\right|\leq\sum_{t=1}^{T}L\epsilon\sum_{\tau=1}^{T}|\alpha^*_{n,\tau}|\leq\epsilon LTC. \tag{74}$$

□

In addition, under the kernel bound in (a3) and uniform convergence in (69) which implies $\sup_{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{X}}\mathbf{z}_n^\top(\mathbf{x}_\tau)\mathbf{z}_n(\mathbf{x}_t)\leq 1+\epsilon$ holds with probability at least $1-2^8(\frac{\sigma_n}{\epsilon})^2\exp(-\frac{D\epsilon^2}{4d+8})$, it can be written that

$$\|\theta^*_n\|^2\leq\|\sum_{t=1}^{T}\alpha^*_{n,t}\mathbf{z}_n(\mathbf{x}_t)\|^2\leq|\sum_{t=1}^{T}\sum_{\tau=1}^{T}\alpha^*_{n,t}\alpha^*_{n,\tau}\mathbf{z}_n^\top(\mathbf{x}_t)\mathbf{z}_n(\mathbf{x}_\tau)|\leq(1+\epsilon)C^2. \tag{75}$$

Combining Lemma 1 with Lemma 5 and (75), it can be concluded that the following bound holds with probability at least $1-2^8(\frac{\sigma_n}{\epsilon})^2\exp(-\frac{D\epsilon^2}{4d+8})$,

$$\begin{aligned}&\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t),y_t)]-\sum_{t=1}^{T}\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\\&=\sum_{t=1}^{T}\mathbb{E}[\mathcal{L}(\hat{f}_t(\mathbf{x}_t),y_t)]-\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)+\sum_{t=1}^{T}\mathcal{L}(\hat{f}^*_{t,n}(\mathbf{x}_t),y_t)-\sum_{t=1}^{T}\mathcal{L}(f^*_n(\mathbf{x}_t),y_t)\\&<\frac{2^b}{\eta}\ln N+\frac{NJ(1+\epsilon)C^2}{2\eta\eta_e^2}+\frac{\eta L^2T}{2}+\eta_eJT+\epsilon LTC+\frac{\eta N^2JT}{2^{b+1}\eta_e^2}\end{aligned} \tag{76}$$

which completes the proof of Theorem 2.

C. Relationship between OMKL-GF and Raker

In this section, we compare our proposed OMKL-GF with Raker (Shen et al., 2019) presented in Algorithm 3.

Algorithm 3 Raker (Shen et al., 2019)

Input: Kernels κ_n, $n=1,...,N$, step size $\eta>0$, and the number of features D.
Initialize: $\boldsymbol{\theta}_{n,1}=\mathbf{0}$, $w_{n,1}=1$, $n=1,...,N$
for $t=1,...,T$ **do**
  Receive one datum $\mathbf{x}_t$.
  Construct $\mathbf{z}_n(\mathbf{x}_t)$ via (5) for $n=1,\ldots,N$.
  Predict $\hat{f}_t(\mathbf{x}_t)=\sum_{n=1}^{N}\frac{w_{n,t}}{\sum_{m=1}^{N}w_{m,t}}\hat{f}_{\text{RF},n}(\mathbf{x}_t)$ with $\hat{f}_{\text{RF},n}(\mathbf{x}_t)$ in (8).
  for $n=1,\ldots,N$ **do**
    Obtain loss $\mathcal{L}(\hat{f}_{RF,n}(\mathbf{x}_t),y_t)$.
    Update $\boldsymbol{\theta}_{n,t+1}$ via (12).
    Update $w_{n,t+1}$ via (13).
  end for
end for

Note that both OMKL-GF and Raker utilizes random feature approximation to make the kernel-based learning task scalable. While Raker employs all kernels in the dictionary for function approximation at each time instance, our proposed OMKL-GF chooses a *time-varying* subset of kernels at each time instant by adaptively pruning irrelevant kernels. Experiments show that OMKL-GF can attain lower MSE and execution time in comparison with Raker by *actively* choosing a subset of kernels.