Variance Reduced Online Gradient Descent for Kernelized Pairwise Learning with Limited Memory Supplementary Material

Hilal AlQuabeh HILAL.ALQUABEH@MBZUAI.AC.AE
Bhaskar Mukhoty BHASKAR.MUKHOTY@GMAIL.COM
Bin Gu BIN.GU@MBZUAI.AC.AE
Machine Learning Department, Mohamed bin Zayed University of Artificial Intelligence

Appendix A. Proofs

Assumption 1 (M-smoothness) *Assume for any $a \in \mathcal{Z}^2 \times \mathcal{H}$, the gradient of the loss function $\nabla\ell(a)$ is M-Lipschitz continuous, i.e. $\forall w, w' \in \mathcal{H}$,*

$$\|\nabla\ell(a) - \nabla\ell(a')\| \leq M \left\|a - a'\right\|_2 .$$

Assumption 2 (Convexity) *Assume for any $z, z' \in \mathcal{Z}$, the loss function $\ell(\cdot, z, z')$ is convex function, i.e. $\forall w, w' \in \mathcal{H}$,*

$$\ell(w, z, z') \geq \ell(w', z, z') + \nabla\ell(w', z, z')^T(w - w').$$

Assumption 3 (Finite Kernel) *Assume for any ρ-probability measure on $\mathcal{X}^2$ the positive kernel function $k : \mathcal{X}^2 \times \mathcal{X}^2 \to \mathbb{R}$ is ρ-integrable, i.e. for any $(x, x') \in \mathcal{X}^2$,*

$$\int\int_{\mathcal{X}^2} k((x, x'), (\hat{x}, \hat{x}'))d\rho(\hat{x})d\rho(\hat{x}') < \infty.$$

A.1. Proof of Lemma 1

Proof Let x_i sampled example from the history of examples, where $i \sim uniform[1, t-1]$, then since the loss function is M-smooth we have,

$$\|\nabla\ell(w, z_t, (x, y)) - \nabla\ell(w, z_t, (\mathbb{E}x, y))\|_2 \leq M \|x - \mathbb{E}x\|_2$$

then by adding and subtracting $\mathbb{E}\nabla\ell(w, z_t, z)$ to LHS after squaring both sides, and denote $\mathbb{E}z = (\mathbb{E}x, y)$ we have,

$$\begin{aligned}
&\Leftrightarrow \|\nabla\ell(w, z_t, z) - \mathbb{E}\nabla\ell(w, z_t, z) + \mathbb{E}\nabla\ell(w, z_t, z) - \nabla\ell(w, z_t, \mathbb{E}z)\|^2 \leq M^2\|x - \mathbb{E}x\|^2 \\
&\Leftrightarrow \|\nabla\ell(w, z_t, z) - \mathbb{E}\nabla\ell(w, z_t, z)\|^2 + \|\mathbb{E}\nabla\ell(w, z_t, z) - \nabla\ell(w, z_t, \mathbb{E}z)\|^2 \\
&\quad - 2(\nabla\ell(w, z_t, z) - \mathbb{E}\nabla\ell(w, z_t, z))^T(\mathbb{E}\nabla\ell(w, z_t, z) - \nabla\ell(w, z_t, \mathbb{E}z)) \leq M^2\|x - Ex\|^2
\end{aligned} \tag{1}$$

taking expectation on both sides w.r.t. uniform distribution of i, and rearrange to have,

$$\begin{aligned}
&\mathbb{E}\|\nabla\ell(w,z_t,z)-\mathbb{E}\nabla\ell(w,z_t,z)\|^2+\mathbb{E}\|\mathbb{E}\nabla\ell(w,z_t,z)-\nabla\ell(w,z_t,\mathbb{E}z)\|^2\\
&\quad-2\underbrace{\mathbb{E}(\nabla\ell(w,z_t,z)-\mathbb{E}\nabla\ell(w,z_t,z))^T}_{=0}(\mathbb{E}\nabla\ell(w,z_t,z)-\nabla\ell(w,z_t,\mathbb{E}z))\leq M^2\mathbb{E}\|x-Ex\|^2\\
\Leftrightarrow&\mathbb{E}\|\nabla\ell(w,z_t,z)-\mathbb{E}\nabla\ell(w,z_t,z)\|^2\leq-\|\mathbb{E}\nabla\ell(w,z_t,z)-\nabla\ell(w,z_t,\mathbb{E}z)\|^2+M^2\mathbb{E}\|x-Ex\|^2\\
&\qquad\leq M^2\mathbb{E}\|x-Ex\|^2 \qquad (2)
\end{aligned}$$

Recall the definition of the variance of stochastic gradient to have final results. This completes the proof. ■

A.2. Proof of Lemma 2

Proof Let $g_t(\cdot)=\frac{1}{|B_t|}\sum_{j\in B_t}\ell(\cdot,z_t,z_j)$ be convex function for all $t\geq 1$ where B_t is the buffer of uniformly sampled *i.i.d.* history examples. Let $u_t\in\partial g_t(w_{t-1})$. If we take the distance of two subsequent models to the optimal model we have,

$$\begin{aligned}
&\|w_t-\bar{w}^*\|^2-\|w_{t-1}-\bar{w}^*\|^2=\|w_{t-1}-\eta_t u_t-\bar{w}^*\|^2-\|w_{t-1}-\bar{w}^*\|^2\\
&=\|w_{t-1}-\bar{w}^*\|^2-2\eta_t u_t^T(w_{t-1}-\bar{w}^*)+\eta_t^2\|u_t\|^2-\|w_{t-1}-\bar{w}^*\|^2\\
&=-2\eta_t u_t^T(w_{t-1}-\bar{w}^*)+\eta_t^2\|u_t\|^2\\
&\leq-2\eta(g_t(w_{t-1})-g_t(\bar{w}^*))+\eta_t^2\|u_t\|^2 \qquad (3)
\end{aligned}$$

Where the last inequality implements Assumption 2, i.e. $-u_t^T(w_{t-1}-\bar{w}^*)\leq-(g_t(w_{t-1})-g_t(\bar{w}^*))$.

Setting the step size $\eta_t=\eta$ for all t, and take the expectation w.r.t. the uniform randomness of the history points, and assume that if w is fixed then $\mathbb{E}g_t(\cdot)=L_t(\cdot)$

$$L(w_{t-1})-L(\bar{w}^*)\leq\frac{\|w_{t-1}-\bar{w}^*\|^2-\|w_t-\bar{w}^*\|^2}{2\eta}+\frac{\eta\mathbb{E}\|u_t\|^2}{2} \qquad (4)$$

Finally using the identity $\mathbb{E}\|u_t\|^2=\mathbb{V}(u_t)+\|\mathbb{E}u_t\|^2$, summing from $t=2$ to $t=T$ and setting $w_1=0$, would completes the proof. ■

A.3. Proof of Theorem 5

Proof

Starting from equation 4 with fact that the cluster-based buffer loss is unbiased of true local loss;

$$L(w_{t-1})-L(\bar{w}^*)\leq\frac{\|w_{t-1}-\bar{w}^*\|^2-\|w_t-\bar{w}^*\|^2}{2\eta}+\frac{\eta\mathbb{E}\|u_t\|^2}{2}$$

and using the M-smoothness of the function g i.e. $L(w_t) \leq L(w_{t-1}) + \mathbb{E}u_t^T(w_t - w_{t-1}) + \frac{M}{2}\|w_t - w_{t-1}\|^2$, and the update $w_t = w_{t-1} - \eta_t v_t$,

$$\mathbb{E}L(w_t) \leq \mathbb{E}L(w_{t-1}) - \eta\|\mathbb{E}u_t\|^2 + \frac{\eta^2 M}{2}\mathbb{E}\|u_t\|^2 \tag{5}$$

By combining the last two inequalities and considering that the expectation is with respect to uniform sampling, we obtain the following inequality.

$$L(w_t) - L(\bar{w}^*) \leq \frac{\|w_{t-1} - \bar{w}^*\|^2 - \|w_t - \bar{w}^*\|^2}{2\eta} + (\frac{\eta}{2} + \frac{\eta^2 M}{2})\mathbb{E}\|u_t\|^2 - \eta\|\mathbb{E}u_t\|^2$$

Using the fact that $\mathbb{V}(u_t) = \mathbb{E}\|u_t\|^2 - \|\mathbb{E}u_t\|^2$ and choosing $\eta = (0, \frac{1}{M}]$, we have,

$$L(w_t) - L(\bar{w}^*) \leq \frac{\|w_{t-1} - \bar{w}^*\|^2 - \|w_t - \bar{w}^*\|^2}{2\eta} + \eta\mathbb{V}(u_t)$$

Finally summing from $t = 2$ to $t = T$ and setting $w_1 = 0$ would complete the proof. ■

It is worth noting that our analysis remains valid in both scenarios: the FIFO buffer update, which requires independent examples, and the randomized update of the buffer which doesn't require online independent examples. While there is a coupling between the model w_t and the buffer B_t, as the model w_{t-1} incorporates information from the buffer at the previous step, we can still maintain the validity of the analysis by considering that this coupling is limited, as demonstrated in previous research (e.g., Zhao et al. (2011)). Although the gradient is not an unbiased statistic due to this coupling, we argue that the impact on the analysis is minimal.

Moreover, it is important to highlight that the main difference lies in the buffer size. In the case of coupling, the buffer size needs to be at least $\log T$, as determined through rigorous analysis utilizing techniques such as Rademacher complexity or covering number (for more details, refer to Kar et al. (2013) and Wang et al. (2012)). These analyses provide a deeper understanding of the underlying mechanisms and further support the validity of our approach.

A.4. Certificate of Variance Reduction

Assume that there exist κ_t clusters, and denote $\hat{u}_t(\cdot)$ the cluster-based buffer gradient constructed using online stratified sampling 1, and $u_t(\cdot)$ represents the estimate obtained from uniform sampling without online clustering,

$$
\begin{aligned}
\mathbb{V}(u_t) &= \frac{1}{\kappa_t}\mathbb{E}_i\|\nabla\ell(w,z_t,z_i) - \mathbb{E}_i\nabla\ell(w,z_t,z)\|^2 \\
&= \frac{1}{\kappa_t}\mathbb{E}_i\|\nabla\ell(w,z_t,z_i)\|^2 - \|\mathbb{E}_i\nabla\ell(w,z_t,z)\|^2 \\
&\overset{(a)}{=} \frac{1}{\kappa_t}\mathbb{E}_{\mathcal{C}_j^t}\mathbb{E}_{i|i\in\mathcal{C}_j^t}\|\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t)\|^2 - \|\mathbb{E}_i\nabla\ell(w,z_t,z)\|^2 \\
&= \frac{1}{\kappa_t}\sum_{j=1}^{\kappa_t}\frac{c_j^t}{t-1}\mathbb{E}_{i|i\in\mathcal{C}_j^t}\|\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t)\|^2 - \|\mathbb{E}_i\nabla\ell(w,z_t,z)\|^2 \\
&= \frac{1}{\kappa_t}\sum_{j=1}^{\kappa_t}\frac{c_j^t}{t-1}\mathbb{V}_{i|i\in\mathcal{C}_j^t}\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t) + \|\mathbb{E}_{i|i\in\mathcal{C}_j^t}\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t)\|^2 - \|\mathbb{E}_i\nabla\ell(w,z_t,z)\|^2 \\
&\overset{(b)}{=} \frac{1}{\kappa_t}\sum_{j=1}^{\kappa_t}\frac{c_j^t}{t-1}\mathbb{V}_{i|i\in\mathcal{C}_j^t}\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t) + \|\mathbb{E}_{i|i\in\mathcal{C}_j^t}\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t) - \mathbb{E}_i\nabla\ell(w,z_t,z)\|^2 \\
&\geq \frac{1}{\kappa_t}\sum_{j=1}^{\kappa_t}\frac{c_j^t}{t-1}\mathbb{V}_{i|i\in\mathcal{C}_j^t}\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t) = \mathbb{V}(\hat{u}_t) \qquad (6)
\end{aligned}
$$

where equality (a) and equality (b) implements total expectation, i.e. $\mathbb{E}_{\mathcal{C}_j^t}\mathbb{E}_{i|i\in\mathcal{C}_j^t}\nabla\ell(w,z_t,z_i|i\in\mathcal{C}_j^t) = \mathbb{E}_i\nabla\ell(w,z_t,z_i)$. The reduction in variance is influenced by the variances within each partition and the number of examples in it. If each cluster has same number of examples $(t-1)/\kappa_t$, we can observe that the bound becomes $1/\kappa_t$. Note that the maximum reduction in variance is $(t-1)$, which is the case of full gradient. It is worth noting that the variance reduction assumes comparable variances among clusters. If the clusters have different variances, it is advisable to sample more from the high-variance cluster. This extension can be easily incorporated into our algorithm by considering the running variances of each cluster.

A.5. Proof of Theorem 8

The study in Bach (2017) assumes that the true and approximated kernel functions belong to $L^2(\rho)$ i.e. space of square integrable functions (under the assumption 3), with the space $\mathcal{H}$ being dense in $L^2(\rho)$. Before proving the theorem, the following Corollary bounds the error of the pairwise kernel using main theorem in Rahimi and Recht (2007).

Corollary 1 *Given $x_1, x_2, x_1', x_2' \in \mathcal{X}$, and pairwise kernel k defined on $\mathcal{X}^2 \times \mathcal{X}^2$, the random Fourier estimation of the kernel has mistake bounded with probability at least $1 - 8\exp\frac{-D^2\delta}{2}$ as follow,*

$$|\hat{k}_{(x_1,x_2)}(x_1', x_2') - k_{(x_1,x_2)}(x_1', x_2')| \leq \delta$$

The proof follows from claim 1 in Rahimi and Recht (2007) and the definitions of pairwise kernel k, which has four sources of errors.

Proof

$$\begin{aligned}\sum_{t=2}^{T} L_t(\bar{w}^*) - \sum_{t=2}^{T} L_t(w^*) &= \sum_{t=2}^{T} \frac{1}{t-1} \sum_{i=1}^{t-1} \ell(\bar{w}^*, z_t, z_i) - \sum_{t=2}^{T} \frac{1}{t-1} \sum_{i=1}^{t-1} \ell(w^*, z_t, z_i) \\ &= \sum_{t=2}^{T} \frac{1}{t-1} \sum_{i=1}^{t-1} \ell(\bar{w}^*, z_t, z_i) - \ell(w^*, z_t, z_i) \\ &\leq \sum_{t=2}^{T} \frac{1}{t-1} \sum_{i=1}^{t-1} \frac{M}{2} \|\bar{w}^* - w^*\|_2^2 \end{aligned} \tag{7}$$

where last inequality applies assumption 1, and that fact that $\nabla \ell(w^*, z, z') = 0$ for any $z, z' \in \mathcal{Z}$. Using the Representer theorem and the fact that the space $\mathcal{H}$ and $\hat{\mathcal{H}}$ are dense in $L^2(\rho)$ (space of squared integrable function, under assumption 3). Hence, we can approximate any function in $\mathcal{H}$ by a function in $L^2(\rho)$, i.e. without loss of generality we assume that $\bar{w}^* = \sum_{j=1,k\neq j}^{t-1} \alpha^*_{j,k} \hat{k}_{z_t,z_i}$, then we have $\|\bar{w}^* - w^*\|_2^2 \leq \| \sum_{j=1,k\neq j}^{t-1} \alpha^*_{j,k} (\hat{k}_{z_t,z_i} - k_{z_t,z_i})\|^2_{L^2(\rho)}$ using the fact that $\|\cdot\|_{L^2(\rho)} \geq \|\cdot\|_2$, finally using the triangle inequality we have,

$$\begin{aligned}\sum_{t=2}^{T} L_t(\bar{w}^*) - \sum_{t=2}^{T} L_t(w^*) &\leq \sum_{t=2}^{T} \sup_{x_i, x_t \in \mathcal{X}} \frac{M}{2} \sum_{j=1,k\neq j}^{t-1} \|\alpha^*_{j,k} (\hat{k}_{z_t,z_i} - k_{z_t,z_i})\|^2_{L^2(\rho)} \\ &\overset{a}{\leq} \sum_{t=2}^{T} \frac{M}{2} \sum_{j=1,k\neq j}^{t-1} (\alpha^*_{j,k})^2 \delta^2 \\ &\overset{b}{\leq} \sum_{t=2}^{T} \frac{M}{2} \delta^2 \Big(\sum_{j=1,k\neq j}^{t-1} |(\alpha^*_{j,k})| \Big)^2 \\ &= \frac{M}{2} T \|w^*\|_1^2 \delta^2 \end{aligned} \tag{8}$$

where inequality (a) implements corollary 1, inequality (b) use the fact that sum of squares is less than the square of sum, and last equality assumes $\|w^*\|_1 = \sum_{i,j\neq i}^{T} |a^*_{i,j}|$. This completes the proof. ■

Appendix B. More Experiemnts

B.1. Number of Clusters

The number of clusters in our algorithm is determined by the hyperparameter epsilon. In our experiments, we set a maximum limit on the cluster number, even if epsilon allows for more clusters. This limit helps to control memory requirements and computational costs.

In the figure below, we illustrate the impact of different cluster limits on the AUC score for the "a9a" dataset, using a small epsilon value. The results demonstrate that as the number of clusters increases, the AUC score takes more time to reach its maximum value. However, it is also evident that by limiting the number of clusters, the variance is significantly reduced while still achieving satisfactory performance.

Table 1: Datasets used in the experiments, where N_-/N_+ is the ratio of negative to positive examples.

Dataset	Size	Features	N_-/N_+
diabetes	768	8	34.90
ijcnn1	141,691	22	9.45
a9a	32,561	123	3.15
MNIST	60,000	784	1.0
covtype	581012	54	1.0
rcv1.binary	20,242	47,236	0.93
usps	9,298	256	1.0
german	20,242	24	2.3

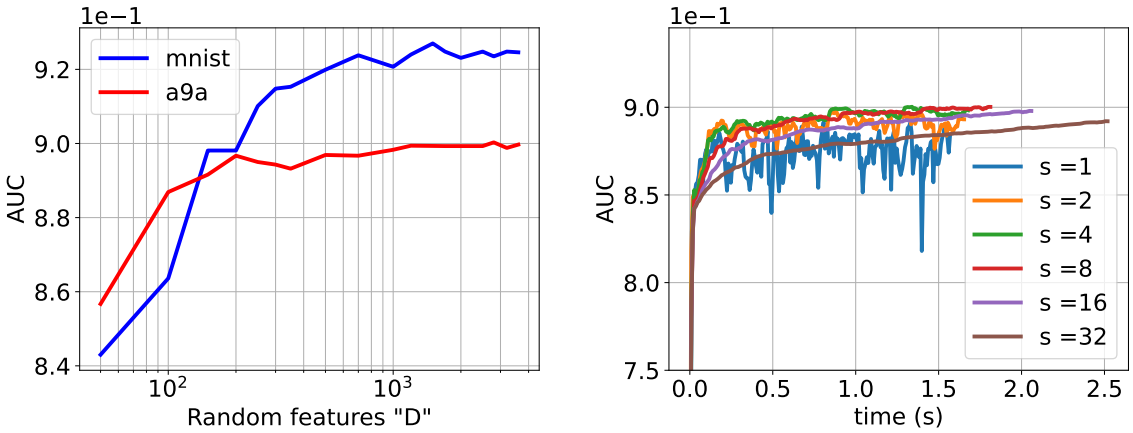

Figure 1: On left, AUC versus number of random features D used to approximate the kernel, for a9a and MNIST datasets. On right,AUC versus maximum number of clusters "s" in algorithm 1 for the dataset "a9a".

B.2. Number of Random Feature

By utilizing the Mercer decomposition theorem and the properties of eigenvalues, we can derive bounds on the number of random Fourier features needed for different decay rates of the eigenvalues. Specifically, for a decay rate of $1/i$, the sufficient number of features is $D \geq 5T \log 2T$. For a decay rate of R^2/i^{2c}, the sufficient number of features is $D \geq T^{1/2c} \log T$. And for a geometric decay rate of r^i $(r > 1)$, the sufficient number of features is $D \geq \log^2 T$.

In our experiments, we focus on a Gaussian kernel with a constant width of $1/d$, where d is the dimension of the input space. For this kernel, the required number of random features is $O(\sqrt{T}\log(T))$. Figure 1 illustrates the results of our experiments, which show that $D = O(\sqrt{T}\log(T))$ is sufficient for a good approximation of the kernel, as the AUC does not improve significantly beyond this point.

B.3. More datasets

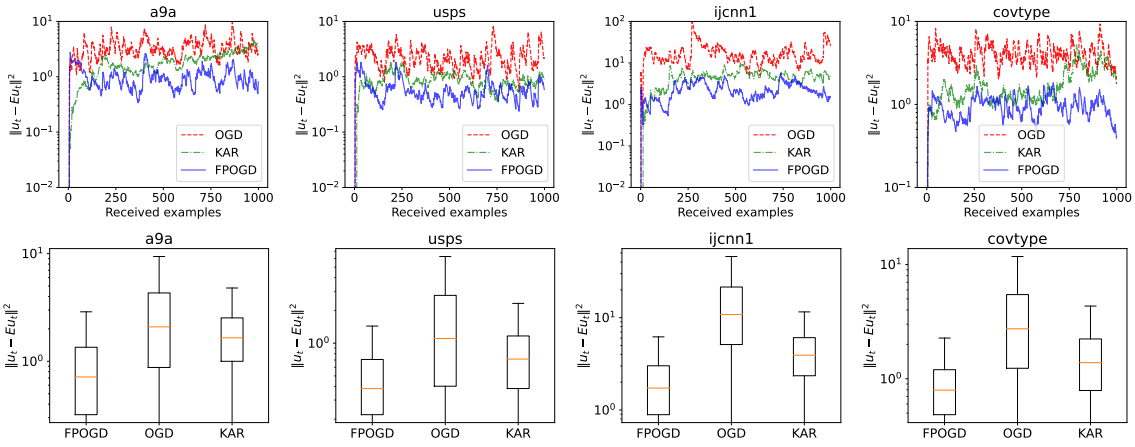

Figure 2: Analyzing Gradient Variance with 4-Size Buffer Across Four Datasets: This study investigates gradient variance using a 4-size buffer across four datasets. The top row shows running variances, while the bottom row presents logarithmic box plots. "Received examples" are online inputs at each time step, with u_t as the online stochastic gradient, and Eu_t as the expected gradient. The red line in the box plots represents the mean variance of stochastic gradients across all algorithms.

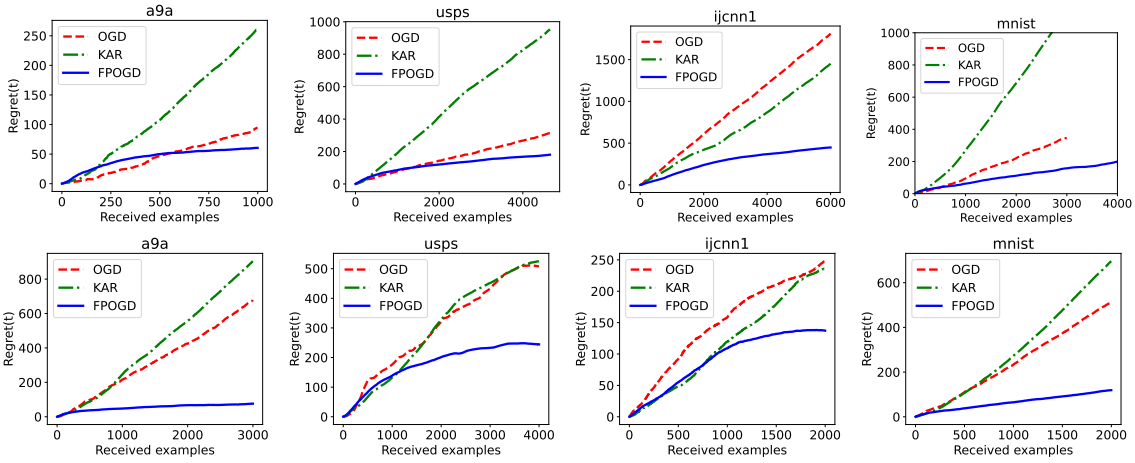

Figure 3: Comparing Regret Across Algorithms and Datasets: This figure illustrates regret, defined as $\sum_t L_t(w_t) - L_t(w^*)$, across diverse algorithms and datasets. The first row depict various algorithms applied to i.i.d. datasets, while the second row features non-i.i.d. datasets generated by sorting examples based on a single feature. Notably, our approach exhibits stronger sublinear regret in the non-i.i.d. setting compared to other algorithms.

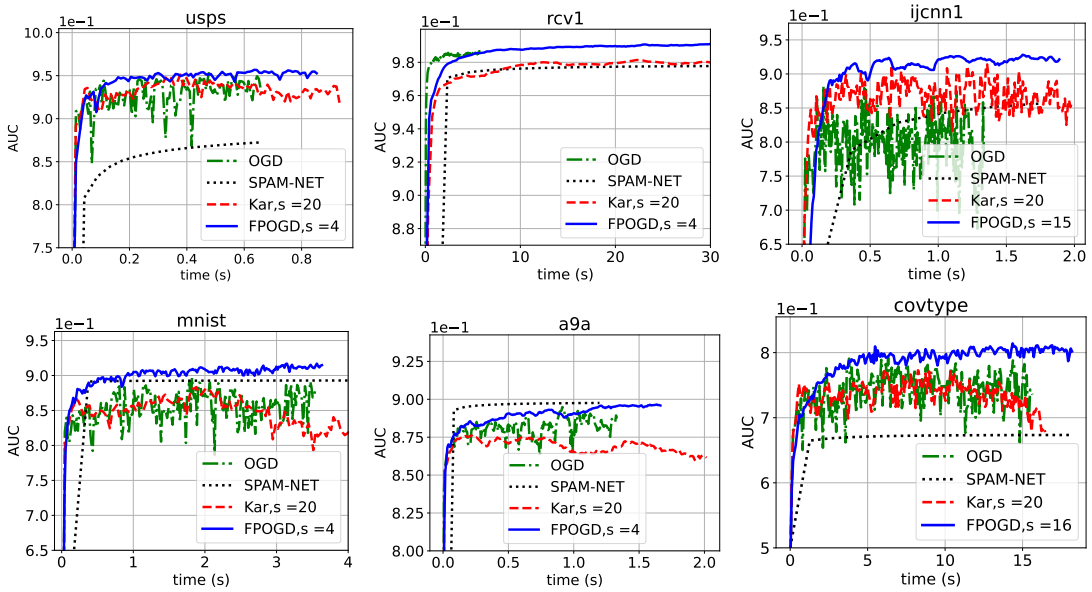

Figure 4: The AUC vs. time comparison of the algorithms in different datasets showing superior performance of the proposed method.

References

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 441–449, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/kar13.html.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Yuyang Wang, Roni Khardon, Dmitry Pechyony, and Rosie Jones. Generalization bounds for online learning algorithms with pairwise loss functions. In *Conference on Learning Theory*, pages 13–1. JMLR Workshop and Conference Proceedings, 2012.

Peilin Zhao, Steven CH Hoi, Rong Jin, and Tianbo YANG. Online auc maximization. *Proceedings of the 28th International Conference on Machine Learning ICML 2011:*, 2011.