Supplementary Material: Online Kernel Learning with a Near Optimal Sparsity Bound

Lijun Zhang ZHANGLIJ@MSU.EDU
Jinfeng Yi YIJINFEN@MSU.EDU
Rong Jin RONGJIN@CSE.MSU.EDU
Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824, USA

Ming Lin LIN-M08@MAILS.TSINGHUA.EDU.CN
Department of Automation, Tsinghua University, Beijing 100084, China

Xiaofei He XIAOFEIHE@CAD.ZJU.EDU.CN
State Key Laboratory of CAD&CG, College of Computer Science, Zhejiang University, Hangzhou 310027, China

A. Proof of Theorem 2

We here prove a lower bound on the number of support vectors to achieve the optimal regret bound.

First, we construct a set of n examples $\mathcal{T}_1 = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where $\langle \kappa(\mathbf{x}_i, \cdot), \kappa(\mathbf{x}_j, \cdot)\rangle_{\mathcal{H}_\kappa} = \delta_{ij}$ and $y_i \in \{1, -1\}$. To make the construction, consider the degree-d polynomial kernel $\kappa(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y})^d$ and an Euclidean space $\mathbb{R}^m$ where $m > n$. Since $m > n$, we can find a set of orthonormal vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ in $\mathbb{R}^m$ such that $\mathbf{x}_i^T\mathbf{x}_j = 0$ when $i \neq j$ and $\mathbf{x}_i^T\mathbf{x}_i = 1$. It is easy to verify that this construction satisfies our assumption $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \delta_{ij}$. For the Gaussian kernel, when the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ are large enough, we also have $\kappa(\mathbf{x}_i, \mathbf{x}_j) \approx \delta_{ij}$.

Based on $\mathcal{T}_1$, we construct another set $\mathcal{T}_2$: $(\mathbf{z}, u) \in \mathcal{T}_2$ if there exist an index $j \in [n]$ and a function $\xi \in \mathcal{H}_\kappa$ such that

$$\kappa(\mathbf{z}, \cdot) = \kappa(\mathbf{x}_j, \cdot) + \xi, \ u = y_j, \ \text{and} \ \langle \xi, \kappa(\mathbf{x}_i, \cdot)\rangle_{\mathcal{H}_\kappa} = 0, \forall i \in [n]. \tag{13}$$

Thus, for each $(\mathbf{z}, u) \in \mathcal{T}_2$, there is a corresponding $(\mathbf{x}_j, y_j) \in \mathcal{T}_1$ such that the relationships in (13) hold. The existence of $\mathcal{T}_2$ can be proved in a similar way as that of $\mathcal{T}_1$.

Second, we select T distinct training examples $(\mathbf{z}_1, u_1), \ldots, (\mathbf{z}_T, u_T)$ from $\mathcal{T}_2$ such that, for each $(\mathbf{x}, y) \in \mathcal{T}_1$ there are T/n examples constructed from it. Taking logit loss $\ell(y, z) = \ln(1 + \exp(-yz))$ as an example. From the above constructions, it is easy to check that

$$f_* = \frac{R}{\sqrt{n}} \sum_{i=1}^n y_i \kappa(\mathbf{x}_i, \cdot)$$

minimizes the cumulative loss on the T training examples, i.e.,

$$f_* = \underset{\|f\|_{\mathcal{H}_\kappa} \leq R}{\operatorname{argmin}} \sum_{i=1}^T \ell(u_i, f(\mathbf{z}_i)) = \underset{\|f\|_{\mathcal{H}_\kappa} \leq R}{\operatorname{argmin}} \sum_{i=1}^T \ell(u_i, \langle f, \kappa(\mathbf{z}_i, \cdot)\rangle_{\mathcal{H}_\kappa}) = \underset{\|f\|_{\mathcal{H}_\kappa} \leq R}{\operatorname{argmin}} \sum_{j=1}^n \sum_{k=1}^{T/n} \ell(y_j, \langle f, \kappa(\mathbf{x}_j, \cdot) + \xi_{jk}\rangle_{\mathcal{H}_\kappa}),$$

and the minimal loss is given by

$$\epsilon = T \ln(1 + \exp(-R/\sqrt{n})).$$

Choosing

$$R = \sqrt{n} \ln \frac{T}{n},$$

we have

$$\epsilon \leq T\exp(-R/\sqrt{n}) = n.$$

Thus, the optimal solution f_* has n support vectors and the associated loss is $O(n)$.

Third, we examine the performance of Algorithm 1. From Theorem 1, we know that both the regret and the number of support vectors of OSKL is on the order of

$$O(\epsilon + R^2) = O(n[\ln T]^2).$$

Finally, we consider any algorithm that outputs a sequence of kernel classifiers $f_1', \ldots, f_T'$ with no more than $n-1$ support vectors. By our construction, this algorithm must misclassify at least T/n training examples, and the cumulative loss $\sum_{i=1}^T \ell(u_i, f_i'(\mathbf{z}_i))$ must be *larger than*

$$\frac{T}{n}\ln 2 = \Omega\left(\frac{T}{n}\right).$$

Recall that the cumulative loss of f_* is $O(n)$. So, the regret of this algorithm is also larger than $\Omega(T/n)$, which is significantly worse than $O(n[\ln T]^2)$ for large T.

B. Proof of Lemma 1

We first state the Bernstein's inequality for martingales (Cesa-Bianchi & Lugosi, 2006), which lays the foundation of the main results.

Theorem 3. *(Bernstein's inequality for martingales). Let $X_1, \ldots, X_n$ be a bounded martingale difference sequence with respect to the filtration $\mathcal{F} = (\mathcal{F}_i)_{1\leq i\leq n}$ and with $|X_i| \leq K$. Let*

$$S_i = \sum_{j=1}^{i} X_j$$

be the associated martingale. Denote the sum of the conditional variances by

$$\Sigma_n^2 = \sum_{t=1}^{n} \mathrm{E}\left[X_t^2|\mathcal{F}_{t-1}\right].$$

Then, for all constants t, $\nu > 0$,

$$\Pr\left[\max_{i=1,\ldots,n} S_i > t \text{ and } \Sigma_n^2 \leq \nu\right] \leq \exp\left(-\frac{t^2}{2(\nu + Kt/3)}\right),$$

and therefore,

$$\Pr\left[\max_{i=1,\ldots,n} S_i > \sqrt{2\nu t} + \frac{2}{3}Kt \text{ and } \Sigma_n^2 \leq \nu\right] \leq e^{-t}.$$

Proof. Define martingale difference

$$X_t = GZ_t - |\ell'(y_t, f_t(\mathbf{x}_t))|,$$

and martingale $\Lambda_T = \sum_{t=1}^T X_t$. Define

$$K = \max_t |X_t| \leq G.$$

Define the conditional variance Σ_T^2 as

$$\Sigma_T^2 = \sum_{t=1}^{T} \mathrm{E}_{t-1}\left[(GZ_t - |\ell'(y_t, f_t(\mathbf{x}_t))|)^2\right] \leq \sum_{t=1}^{T} G|\ell'(y_t, f_t(\mathbf{x}_t))| = GA_T.$$

Since $A_T \leq 1$, we have $\Sigma_T^2 \leq \frac{G}{T}$. Following Theorem 3, with probability at least $1-\delta$, we have

$$\Lambda_T = \sum_{t=1}^{T} GZ_t - |\ell'(y_t, f_t(\mathbf{x}_t))| \leq \sqrt{2\frac{G}{T}\ln\frac{1}{\delta}} + \frac{2}{3}G\ln\frac{1}{\delta} \leq G\ln\frac{1}{\delta},$$

where the last inequality follows from the fact $T \geq 18/[G\ln(1/\delta)]$. □

C. Proof of Lemma 2

We use the same definitions of X_t, Λ_T, K and Σ_T^2 in Appendix B. Notice that A_T in the upper bound for Σ_T^2 is a random variable, thus we cannot direct apply Theorem 3. To handle this challenge, we make use of the peeling process described in (Bartlett et al., 2005), and have

$$
\begin{aligned}
& \Pr\left(\Lambda_T \geq 2\sqrt{GA_T\tau} + \frac{2}{3}K\tau\right) \\
=& \Pr\left(\Lambda_T \geq 2\sqrt{GA_T\tau} + \frac{2}{3}K\tau, A_T \leq G_1 T\right) \\
=& \Pr\left(\Lambda_T \geq 2\sqrt{GA_T\tau} + \frac{2}{3}K\tau, \Sigma_T^2 \leq GA_T, A_T \leq G_1 T\right) \\
\leq& \Pr\left(\Lambda_T \geq 2\sqrt{GA_T\tau} + \frac{2}{3}K\tau, \Sigma_T^2 \leq GA_T, A_T \leq \frac{1}{T}\right) \\
& + \sum_{i=1}^{m} \Pr\left(\Lambda_T \geq 2\sqrt{GA_T\tau} + \frac{2}{3}K\tau, \Sigma_T^2 \leq GA_T, \frac{2^{i-1}}{T} < A_T \leq \frac{2^i}{T}\right) \\
\leq& \Pr\left(A_T \leq \frac{1}{T}\right) + \sum_{i=1}^{m} \Pr\left(\Lambda_T \geq \sqrt{2\frac{G2^i}{T}\tau} + \frac{2}{3}K\tau, \Sigma_T^2 \leq \frac{G2^i}{T}\right) \\
\leq& \Pr\left(A_T \leq \frac{1}{T}\right) + me^{-\tau},
\end{aligned}
$$

where $m = \lceil \log_2(G_1 T^2) \rceil$, and the last step follows the Bernstein's inequality for martingales. We complete the proof by setting $\tau = \ln(m/\delta)$, and using the assumption $A_T > 1/T$.