Active Nearest Neighbors in Changing Environments: Supplementary Material

Christopher Berlind CBERLIND@GATECH.EDU
Georgia Institute of Technology, Atlanta, GA, USA

Ruth Urner RUTH.URNER@TUEBINGEN.MPG.DE
Max Planck Institute for Intelligent Systems, Tübingen, Germany

1. Proof of Theorem 1

We adapt the proof (guided exercise) of Theorem 19.5 in (Shalev-Shwartz & Ben-David, 2014) to our setting. As is done there, we use the notation $y \sim p$ to denote drawing from a Bernoulli random variable with mean p. We will employ the following lemmas:

Lemma 1 (Lemma 19.6 in (Shalev-Shwartz & Ben-David, 2014))**.** *Let $C_1, \ldots, C_r$ be a collection of subsets of some domain set, $\mathcal{X}$. Let S be a sequence of m points sampled i.i.d. according to some probability distribution, D over $\mathcal{X}$. Then, for every $k \geq 2$,*

$$\mathop{\mathbb{E}}_{S\sim D^m}\left[\sum_{i:|C_i\cap S|<k} \mathbb{P}[C_i]\right] \leq \frac{2rk}{m}\,.$$

Lemma 2 (Lemma 19.7 in (Shalev-Shwartz & Ben-David, 2014))**.** *Let $k \geq 10$ and let $Z_1, \ldots, Z_k$ be independent Bernoulli random variables with $\mathbb{P}[Z_i = 1] = p_i$. Denote $p = \frac{1}{k}\sum_i p_i$ and $p' = \frac{1}{k}\sum_{i=1}^k Z_i$. Then*

$$\mathop{\mathbb{E}}_{Z_1,\ldots,Z_k} \mathop{\mathbb{P}}_{y\sim p}[y \neq \mathbf{1}[p' > 1/2]] \leq \left(1 + \sqrt{\frac{8}{k}}\right) \mathop{\mathbb{P}}_{y\sim p}[y \neq \mathbf{1}[p > 1/2]]\,.$$

Before we prove the theorem, we show the following:

Claim 1 (Ex. 3 of Chapter 19 in (Shalev-Shwartz & Ben–David, 2014))**.** *Fix some $p, p' \in [0,1]$ and $y' \in \{0,1\}$. Then*

$$\mathop{\mathbb{P}}_{y\sim p}[y \neq y'] \leq \mathop{\mathbb{P}}_{y\sim p'}[y \neq y'] + |p - p'|\,.$$

Proof. If $y' = 0$, we have

$$\begin{aligned}\mathop{\mathbb{P}}_{y\sim p}[y \neq y'] &= p = p - p' + p' \\ &= \mathop{\mathbb{P}}_{y\sim p'}[y \neq y'] + p - p' \\ &\leq \mathop{\mathbb{P}}_{y\sim p'}[y \neq y'] + |p - p'|.\end{aligned}$$

If $y' = 1$, we have

$$\begin{aligned}\mathop{\mathbb{P}}_{y\sim p}[y \neq y'] &= 1 - p = 1 - p - p' + p' \\ &= \mathop{\mathbb{P}}_{y\sim p'}[y \neq y'] - p + p' \\ &\leq \mathop{\mathbb{P}}_{y\sim p'}[y \neq y'] + |p - p'|.\end{aligned}$$

□

Proof of Theorem 1. Let h_{ST} denote the output classifier of Algorithm 1. Let $\mathcal{C} = \{C_1, \ldots, C_r\}$ denote an ϵ-cover of the target support $(\mathcal{X}_T, \rho)$, that is, $\bigcup_i C_i = \mathcal{X}_T$ and each C_i has diameter at most ϵ. Without loss of generality, we assume that the C_i are disjoint and for a domain point $x \in \mathcal{X}$ we let $C(x)$ denote the element of $\mathcal{C}$ that contains x. Let $L = T^l \cup S$ denote the (k,k')-NN-cover of T that ANDA uses (that is, the set of labeled points that h_{ST} uses for prediction). We bound its expected loss as follows:

$$\begin{aligned}&\mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}[\mathcal{L}_{P_T}(h_{ST})] \\ &= \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[h_{ST}(x) \neq y]\right] \\ &\leq \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[h_{ST}(x) \neq y \wedge \rho(x, x_{k'}(x,T)) > \epsilon]\right. \\ &\quad \left. + \mathop{\mathbb{P}}_{(x,y)\sim P_T}[h_{ST}(x) \neq y \wedge \rho(x, x_{k'}(x,T)) \leq \epsilon]\right] \\ &= \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[\rho(x, x_{k'}(x,T)) > \epsilon]\right] \\ &\quad + \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[h_{ST}(x) \neq y \ \wedge \ \rho(x, x_{k'}(x,T)) \leq \epsilon]\right] \\ &= \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[\rho(x, x_{k'}(x,T)) > \epsilon]\right] \\ &\quad + \mathop{\mathbb{E}}_{x\sim D_T}\left[\mathop{\mathbb{P}}_{\substack{y\sim\eta(x)\\ T\sim {D_T}^{m_T}}}[h_{ST}(x) \neq y \ \wedge \ \rho(x, x_{k'}(x,T)) \leq \epsilon]\right],\end{aligned}$$

where the last equality holds by Fubini's theorem. Then we

have

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}[\mathcal{L}_{P_T}(h_{ST})] \\
\leq\ &\mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[\rho(x,x_{k'}(x,T))>\epsilon]\right] \\
&+\mathop{\mathbb{E}}_{x\sim D_T}\left[\mathop{\mathbb{P}}_{\substack{y\sim\eta(x)\\ T\sim {D_T}^{m_T}}}[h_{ST}(x)\neq y\ \mid\ \rho(x,x_{k'}(x,T))\leq\epsilon]\right] \\
\leq\ &\mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[|T\cap C(x)|<k']\right] \\
&+\mathop{\mathbb{E}}_{x\sim D_T}\left[\mathop{\mathbb{P}}_{\substack{y\sim\eta(x)\\ T\sim {D_T}^{m_T}}}[h_{ST}(x)\neq y\ \mid\ \rho(x,x_{k'}(x,T))\leq\epsilon]\right],
\end{aligned}
$$

where for the first summand of the last inequality, we used that a point x can only have distance more than ϵ to its k'-th nearest neighbor in T if $C(x)$ is hit less than k' times by T. Lemma 1 implies that this first summand is bounded in expectation by $\frac{2\,\mathrm{N}_\epsilon(\mathcal{X}_T,\rho)\,k'}{m_T}$.

To bound the second summand, we now first fix a sample T and a point x such that $\rho(x,x_{k'}(x,T))\leq\epsilon$ (and condition on these). Since the set of labeled points $L=T^l\cup S$ used for prediction is an (k,k')-NN-cover of T, Lemma 1 implies that there are at least k labeled points in L at distance at most 3ϵ from x. Let $k(x,L)=\{x_1,\dots,x_k\}$ be the k nearest neighbors of x in L, let $p_i=\eta(x_i)$ and set $p=\frac{1}{k}\sum_i p_i$. Now we get

$$
\begin{aligned}
&\mathop{\mathbb{P}}_{y_1\sim p_1,\dots y_k\sim p_k,y\sim\eta(x)}[h_{ST}(x)\neq y] \\
&=\mathop{\mathbb{E}}_{y_1\sim p_1,\dots y_k\sim p_k}\left[\mathop{\mathbb{P}}_{y\sim\eta(x)}[h_{ST}(x)\neq y]\right] \\
&\leq\mathop{\mathbb{E}}_{y_1\sim p_1,\dots y_k\sim p_k}\left[\mathop{\mathbb{P}}_{y\sim p}[h_{ST}(x)\neq y]\right]+|p-\eta(x)| \\
&\leq\left(1+\sqrt{\frac{8}{k}}\right)\mathop{\mathbb{P}}_{y\sim p}[y\neq\mathbf{1}[p>1/2]]+|p-\eta(x)|,
\end{aligned}
$$

where the first inequality follows from Claim 1 and the second from Lemma 2. We have

$$
\begin{aligned}
\mathop{\mathbb{P}}_{y\sim p}[\mathbf{1}[p>1/2]\neq y]&=p \\
&=\min\{p,1-p\} \\
&\leq\min\{\eta(x),1-\eta(x)\}+|p-\eta(x)|\,.
\end{aligned}
$$

Further, since the regression function η is λ-Lipschitz and $\rho(x_i,x)\leq 3\epsilon$ for all i, we have

$$
\begin{aligned}
|p-\eta(x)|&=\left|\left(\frac{1}{k}\sum_i p_i\right)-\eta(x)\right| \\
&=\left|\left(\frac{1}{k}\sum_i \eta(x_i)\right)-\eta(x)\right| \\
&=\left|\left(\frac{1}{k}\sum_i \eta(x_i)-\eta(x)+\eta(x)\right)-\eta(x)\right| \\
&\leq\left|\left(\frac{1}{k}\sum_i 3\lambda\epsilon+\eta(x)\right)-\eta(x)\right| \\
&=\left|3\lambda\epsilon+\left(\frac{1}{k}\sum_i \eta(x)\right)-\eta(x)\right|\ =\ 3\lambda\epsilon.
\end{aligned}
$$

Thus, we get

$$
\begin{aligned}
&\mathop{\mathbb{P}}_{y_1\sim p_1,\dots y_k\sim p_k,y\sim\eta(x)}[h_{ST}(x)\neq y] \\
&=\mathop{\mathbb{E}}_{y_1\sim p_1,\dots y_k\sim p_k}\left[\mathop{\mathbb{P}}_{y\sim\eta(x)}[h_{ST}(x)\neq y]\right] \\
&\leq\left(1+\sqrt{\frac{8}{k}}\right)\mathop{\mathbb{P}}_{y\sim p}[y\neq\mathbf{1}[p>1/2]]+|p-\eta(x)| \\
&\leq\left(1+\sqrt{\frac{8}{k}}\right)(\min\{\eta(x),1-\eta(x)\}+|p-\eta(x)|)+|p-\eta(x)| \\
&\leq\left(1+\sqrt{\frac{8}{k}}\right)(\min\{\eta(x),1-\eta(x)\})+3|p-\eta(x)| \\
&\leq\left(1+\sqrt{\frac{8}{k}}\right)(\min\{\eta(x),1-\eta(x)\})+9\lambda\epsilon.
\end{aligned}
$$

Since this holds for all samples T and points x with $\rho(x,x_{k'}(x,T))\leq\epsilon$, we get,

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim D_T}\left[\mathop{\mathbb{P}}_{\substack{y\sim\eta(x)\\ T\sim {D_T}^{m_T}}}[h_{ST}(x)\neq y\ \mid\ \rho(x,x_{k'}(x,T))\leq\epsilon]\right] \\
\leq\ &\mathop{\mathbb{E}}_{x\sim D_T}\left[\left(1+\sqrt{\frac{8}{k}}\right)(\min\{\eta(x),1-\eta(x)\})+9\lambda\epsilon\right] \\
=\ &\left(1+\sqrt{\frac{8}{k}}\right)\mathop{\mathbb{E}}_{x\sim D_T}[(\min\{\eta(x),1-\eta(x)\})]+9\lambda\epsilon \\
=\ &\left(1+\sqrt{\frac{8}{k}}\right)\mathcal{L}_T(h_T^*)+9\lambda\epsilon.
\end{aligned}
$$

This yields

$$\begin{aligned}
&\mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}[\mathcal{L}_{P_T}(h_{ST})] \\
&\leq \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[|T\cap C(x)|<k']\right] \\
&+ \mathop{\mathbb{E}}_{T\sim {D_T}^{m_T}}\left[\mathop{\mathbb{P}}_{(x,y)\sim P_T}[h_{ST}(x)\neq y \mid \rho(x,x_{k'}(x,T))\leq\epsilon]\right] \\
&\leq \frac{2\,\mathrm{N}_\epsilon(\mathcal{X}_T,\rho)\,k'}{m_T}+\left(1+\sqrt{\frac{8}{k}}\right)\mathcal{L}_T(h_T^*)+9\lambda\epsilon,
\end{aligned}$$

which completes the proof. □

2. Proof of Corollary 1

Proof. We need to show that for every $\alpha > 0$, there exists an index i_0, such that $\mathbb{E}_{T\sim P_T^{m_i}}[\mathcal{L}_T(\mathrm{ANDA}(S_i,T,k_i,{k'}_i))] = \mathcal{L}_T(h^*)+\alpha$ for all $i\geq i_0$. Let $P_T\in\mathcal{P}(\mathcal{X},\rho)$ and α be given.

Let γ be so that $9\gamma\leq\alpha/3$. Since η is uniformly continuous, there is a δ, such that for all $x,x'\in\mathcal{X}$, $\rho(x,x')\leq\delta\Rightarrow|\eta(x)-\eta(x')|\leq\gamma$. Note that the only way we used the λ-Lipschitzness in the proof of Theorem 1 is by using that for any two points x,x' that lie in a common element C of an ϵ-cover of the space, we have $|\eta(x)-\eta(x')|\leq\lambda\epsilon$. Thus, we could now repeat the proof of Theorem 1, using a δ-cover of the space and obtain that

$$\begin{aligned}
&\mathop{\mathbb{E}}_{T\sim D_T^{m_T}}[\mathcal{L}_T(\mathrm{ANDA}(S,T,k,k'))] \\
&\qquad\leq (1+\sqrt{8/k})\mathcal{L}_T(h^*)+9\gamma+\frac{2\,\mathrm{N}_\delta(\mathcal{X}_T,\rho)\,k'}{m_T}.
\end{aligned}$$

for all $k\geq 10$ and $k'\geq k$. Now let i_1 be so that $\sqrt{\frac{8}{k_i}}\leq\frac{\alpha}{3}$ for all $i\geq i_1$. Note that this implies $\sqrt{\frac{8}{k_i}}\mathcal{L}_T(h^*)\leq\frac{\alpha}{3}$ for all $i\geq i_1$. Since $({k'}_i/m_i)\to 0$ as $i\to\infty$, we can choose i_2 be so that $\frac{2\,\mathrm{N}_\delta(\mathcal{X}_T,\rho)\,{k'}_i}{m_i}\leq\alpha/3$ for all $i\geq i_2$. Together these imply that for all $i\geq i_0:=\max\{i_1,i_2\}$, we have $\mathbb{E}_{T\sim P_T^{m_i}}[\mathcal{L}_T(\mathrm{ANDA}(S_i,T,k_i,{k'}_i))] = \mathcal{L}_T(h^*)+\alpha$ as desired. □

3. Proof of Theorem 3

Proof. Recall that, according to the requirements of Theoren 2, we have $m_T > k' = (C+1)k$ for some k that satisfies $k\geq 9\,(d_{\mathsf{VC}}(\mathcal{B})\ln(2m_T)+\ln(6/\delta))$. Since D_T has a continuous density function, for every point x in $\mathcal{X}_T$ and $0<\epsilon\leq 1$, there is a ball $B^\epsilon(x)$ of target weight exactly ϵ around x (*i.e.* $D_T(B^\epsilon(x))=\epsilon$). For some $w>0$, let $\mathcal{X}_T(\epsilon,w)\subseteq\mathcal{X}_T$ denote the set of points x whose ϵ-ball has weight ratio smaller than w, that is $\mathcal{X}_T(\epsilon,w)=\{x\in\mathcal{X}_T \mid \beta(B^\epsilon(x))<w\}$.

Claim 2.

$$\lim_{w\to 0} D_T(\mathcal{X}_T(\epsilon,w)\cap\mathcal{X}_S)=0$$

Let $\epsilon = Ck/3m_T$. Given the claim (which we prove below), we can choose w small enough such that (with probability at least $1-\delta$), a target sample of size m_T will not hit $\mathcal{X}_T(\epsilon,w)\cap\mathcal{X}_S$. Now we can choose a size M_S for the source sample S large enough such that (with probability $1-2\delta$) ANDA-S will not query any points in $\mathcal{X}_S\setminus\mathcal{X}_T(\epsilon,w)$. This is shown similarly to the proof of Theorem 2 as follows.

First, assume that the sample T is so that the implications of Lemma 2 are satisfied (this also happens with probability at least $(1-\delta)$). Then, by invoking the contrapositive of the first implication in Lemma 2,

$$D_T(B^\epsilon(x))=\epsilon=\frac{Ck}{3m_T}$$

and

$$\frac{Ck}{m_T}\geq\frac{C\,9\,(d_{\mathsf{VC}}(\mathcal{B})\ln(2m_T)+\ln(6/\delta))}{m_T}$$

implies that

$$\widehat{T}(B^\epsilon(x))\leq\frac{Ck}{m_T}.$$

Thus, for all x, the ball $B^\epsilon(x)$ contains at most Ck points from the target sample T.

Now we choose a sufficiently large size for the source sample S, namely

$$m_S\geq M_S=\frac{72\,\ln(6/\delta)m_T}{C\,w}\ln\left(\frac{9\,m_T}{C\,w}\right)$$

for the value of w chosen above. We assume that the sample S is so that the implications of Lemma 2 are satisfied (this, again, holds with probability at least $(1-\delta)$).

Exactly as in the proof of Theorem 2, we can show that, for all x with $\beta(B^\epsilon(x))\geq w$,

$$D_T(B^\epsilon(x))=\frac{Ck}{3m_T}$$

implies

$$\widehat{S}(B^\epsilon(x))\geq\frac{k}{m_S},$$

Thus, for all x with $\beta(B^\epsilon(x))\geq w$, the ball $B^\epsilon(x)$ contains at least k points from the source sample S.

In summary, we have shown that with probability $(1-3\delta)$ over the samples S and T, for all target sample points x, that fall into the source support, we have $\beta(B^\epsilon(x))\geq w$, and for those the ball $B^\epsilon(x)$ contains at most Ck target and at least k source samples points. This implies that for all

target sample points, that fall into the source support, the $k' = (C+1)k$ Nearest Neighbor ball (in $S \cup T$) around x contains at least k points from the source sample and will therefore not be queried.

Proof of Claim 2. Let $(w_i)_{i\in\mathbb{N}}$ be a decreasing sequence that converges to 0. Then the sets $\mathcal{X}_T(\epsilon, w_i)$ are linearly ordered by inclusion (getting smaller as w_i gets smaller). Thus, the limit of the sequence of sets $\mathcal{X}_T(\epsilon, w_i)$ exists and we have

$$\lim_{i\to\infty} \mathcal{X}_T(\epsilon, w_i) = \bigcap_{i=1}^{\infty} \mathcal{X}_T(\epsilon, w_i) \subseteq \mathcal{X}_T \setminus \mathcal{X}_S$$

To see the last inclusion, recall that, by definition, a point x is in the source support $\mathcal{X}_S$ if and only if every ball B around x has positive source mass $D_S(B) > 0$. Hence, in particular $D_S(B^\epsilon(x)) > 0$, which implies that these balls also have strictly positive weight ratio $\beta(B^\epsilon(x)) > 0$. Thus, for every point x in the source support, there exists an i such that $x \notin \mathcal{X}_T(\epsilon, w_i)$, since the w_i converge to 0.

The above set convergence implies

$$\lim_{i\to\infty} D_T(\mathcal{X}_T(\epsilon, w_i)) = D_T(\bigcap_{i=1}^{\infty} \mathcal{X}_T(\epsilon, w_i)) \leq D_T(\mathcal{X}_T \setminus \mathcal{X}_S).$$

This, in turn implies

$$\lim_{i\to\infty} D_T(\mathcal{X}_T(\epsilon, w_i) \cap \mathcal{X}_S) \leq D_T((\mathcal{X}_T \setminus \mathcal{X}_S) \cap \mathcal{X}_S) = 0$$

yielding the claim. □

□

References

Shalev-Shwartz, Shai and Ben-David, Shai. *Understanding Machine Learning*. Cambridge University Press, 2014.