Active Nearest Neighbors in Changing Environments

Christopher Berlind
Georgia Institute of Technology
cberlind@gatech.edu

Ruth Urner
Carnegie Mellon University
rurner@cs.cmu.edu

Abstract

While classic machine learning paradigms assume training and test data are generated from the same process, domain adaptation addresses the more realistic setting in which the learner has large quantities of labeled data from a *source* task but limited or no labeled data from a separate *target* task it is attempting to learn. In this work, we demonstrate that being *active adaptive* yields a way to address the statistical challenges inherent in this setting. We propose a new, non-parametric algorithm, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. We provide both an analysis of finite sample convergence rates of the resulting k-nearest neighbor classifier and an analysis of its querying behavior. In addition, we provide experiments on synthetic data to illustrate the adaptivity and query efficiency of our algorithm.

1 Introduction

In a common model for domain adaptation, the learner receives large amounts of labeled data from a (or several) *source* distribution and unlabeled data from the actual *target* distribution (and possibly a small amount of labeled data from the target task as well). The goal of the learner is to output a good model for the target task. For example, an e-commerce company may want to predict the success of a product in one country when they only have preference data on that product from consumers in a different country. To design methods for this scenario that are statistically consistent with respect to the target task is challenging. This difficulty even occurs in the so-called covariate shift setting, where the change in the environments is restricted to the marginal over the covariates, while the regression functions (the labeling rules) of the involved distributions are identical.

In this work, we demonstrate that being *active adaptive* yields a way to address these challenges. We propose a new, non-parametric algorithm for domain adaptation, ANDA, that combines an active nearest neighbor querying strategy with nearest neighbor prediction. ANDA receives a labeled sample (generated by some source distribution) and an unlabeled sample from the target task and selects a subset of the target data to be labeled. It chooses points from the target to query for labels according to how many source examples lie in a k'-nearest neighbor ball around them (this serves as an indication for how well the area of a point is supported by the source). ANDA then predicts with a k-nearest neighbor classifier on the combined source and target labeled data.

We prove that our algorithm enjoys strong performance guarantees. We provide both an analysis of finite sample convergence rates of the resulting k-nearest neighbor classifier and an analysis of its querying behavior. Remarkably, the predictive quality of the output classifier of ANDA *does not depend on the relatedness of source and target*. ANDA will never suffer a negative transfer. This is in sharp contrast to what we know from non-active domain adaptation methods, that are prone to perform poorly if the source is very different from the target. This robustness is achieved by ANDA adapting its querying behavior to the relatedness of source and target. ANDA will *automatically* make more or less queries to the target sample depending on how well the target is supported by the source, that is depending on whether the source provides sufficiently informative examples or not.

In a bit more detail, our main results are summarized as follows:

Bounding the loss. Theorem 1 provides a finite sample bound on the the expected 0-1 loss of the classifier output by ANDA. This bound depends on the size of the unlabeled target sample and on the Lipschitz constant of the underlying regression function. It does not depend on the size or the generating process of the labeled source sample. In particular, it does not depend on any relatedness measure between the source and target data generating distributions. We also show that, even dropping the Lipschitz condition, ANDA is still consistent with respect to the target task.

Bounding the number of queries. In Theorem 2 we show that, with high probability, ANDA will not query any points that are sufficiently supported by the source data. This implies in particular that, if source and target happen to be identical (or very similar), ANDA will not make any queries at all. Together with the error consistency result, this implies that we get the desired behavior of our active adaptive scheme: The loss of the output classifier always converges to the Bayes optimal while queries are made only in regions where the source is not providing information, that is, where acquiring labels from the target is needed.

Approximation guarantee for (k, k')-NN-cover subproblem. In order to select the subset of target points to be queried, we define a combinatorial optimization problem, the (k, k')-NN-cover problem (Definition 2), that may be of independent interest. We show that it is a special case of the MINIMUM MULTISET MULTICOVER problem. We employ a greedy strategy to find a small (k, k')-NN-cover and argue that this greedy strategy enjoys a $O(\log m)$-approximation guarantee on combined source/target samples of m points.

In addition to the theoretical guarantees, we provide experiments on synthetic data to illustrate the adaptivity and query efficiency of our algorithm.

The idea of incorporating active learning (selective querying strategies) in to the design of algorithms for domain adaptation has recently received some attention in the more application-focused machine learning research community [1, 2, 3]. However, to the best of our knowledge, there has not been any formal analysis of the possibilities of incorporating active learning to facilitate being adaptive to distribution changes. We view our work as a first step in this direction. Please refer to Appendix A for a more comprehensive discussion of related work.

1.1 Notation

Let $(\mathcal{X}, \rho)$ be a separable metric space. We let $B_r(x)$ denote the closed ball of radius r around x. We let $\mathrm{N}_\epsilon(\mathcal{X}, \rho)$ denote the *ϵ-cover-number* of the metric space. That is, the minimum number of subsets $C \subseteq \mathcal{X}$ of diameter at most ϵ that cover the space $\mathcal{X}$.

We consider a binary classification task, where P_S and P_T denote *source* and *target distributions* over $\mathcal{X} \times \{0, 1\}$. We let D_S and D_T denote the source and target marginal distributions over $\mathcal{X}$, respectively. Further, we let $\mathcal{X}_S$ and $\mathcal{X}_T$ denote the *support* of D_S and D_T respectively. That is, for $I \in \{S, T\}$, we have $\mathcal{X}_I := \{x \in \mathcal{X} \ : \ D_I(B_r(x)) > 0 \text{ for all } r > 0\}$. We consider the *covariate shift* setting. That is, the regression function $\eta(x) = \mathbb{P}[y = 1|x]$ is the same for both source and target. We use the notation S and T for i.i.d. samples from P_S and D_T, respectively, and let $|S| = m_S$, $|T| = m_T$, and $m = m_S + m_T$.

For any finite $A \subseteq \mathcal{X}$ and $x \in \mathcal{X}$, the notation $x_1(x, A), \ldots, x_{|A|}(x, A)$ gives an ordering of the elements of A such that $\rho(x_1(x, A), x) \leq \rho(x_2(x, A), x) \leq \cdots \leq \rho(x_{|A|}(x, A), x)$. If A is a labeled sequence of domain points, $A = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$, then we use the same notation for the labels (that is $y_i(x, A)$ denotes the label of the i-th nearest point to x in A). We use the notation $k(x, A) = \{x_1(x, A), \ldots, x_k(x, A)\}$ to denote the set of the k nearest neighbors of x in A.

We are interested in bounding the target loss of a k-nearest neighbor classifier. For a sequence A of labeled points $A = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$ we let h^k_A denote the *k-NN classifier* on A:

$$h^k_A(x) := \mathbf{1}\left[1/k\Sigma^k_{i=1} y_i(x, A) \geq 1/2\right],$$

where $\mathbf{1}[\cdot]$ denotes the indicator function. We denote the *Bayes classifier* by $h^*(x) = \mathbf{1}[\eta(x) \geq 1/2]$ and the *target loss* of a classifier $h : \mathcal{X} \to \{0, 1\}$ by $\mathcal{L}_T(h) = \mathbb{P}_{(x,y)\sim P_T}[y \neq h(x)]$. For a subset $A \subseteq \mathcal{X}$ of the domain that is measurable both with respect to D_S and D_T and satisfies $D_T(A) > 0$, we define the *weight ratio* of A as $\beta(A) := D_S(A)/D_T(A)$. For a collection of subsets $\mathcal{B} \subseteq 2^{\mathcal{X}}$ (for example all balls in $(\mathcal{X}, \rho)$), we let $\mathrm{VCdim}(\mathcal{B})$ denote its VC-dimension.

Algorithm 1 ANDA: Active Nearest Neighbor Domain Adaptation

input Labeled sample S, unlabeled sample T, parameters k, k'
Find $T^l \subseteq T$ such that $S \cup T^l$ is a (k, k')-NN-cover of T
Query the labels of points in T^l
return $h^k_{S \cup T^l}$, the k-NN classifier on $S \cup T^l$

Algorithm 2 Safe: Find (k, k')-NN-cover

input Labeled sample S, unlabeled sample T, parameters k, k'
return $\{x \in T : |k'(x, S \cup T) \cap S| < k\}$

2 The algorithm

In brief, our algorithm receives a labeled sample S (from the source distribution), an unlabeled sample T (from the target distribution), and two parameters k and k'. It then chooses a subset $T^l \subset T$ to be labeled, queries the labels of points in T^l, and outputs a k-NN predictor on $S \cup T^l$ (see Algorithm 1). The subset T^l is chosen so that the resulting labeled set $S \cup T^l$ is a (k, k')-NN-*cover* for the target sample T.

Definition ((k, k')-NN-cover)**.** Let $T \subseteq \mathcal{X}$ be a set of elements in a metric space $(\mathcal{X}, \rho)$ and let $k, k' \in \mathbb{N}$ with $k \leq k'$. A set R is a (k, k')-NN-cover for T, if for every $x \in T$, either $x \in R$ or there are k elements from R among the k' nearest neighbors of x in $T \cup R$, that is $|k'(x, T \cup R) \cap R| \geq k$.

Our loss bound in Section 3 (Theorem 1) holds whenever $T^l \cup S$ is some (k, k')-NN-cover of T. Algorithm 2 provides a simple strategy to find such a cover: add to T^l all points x whose k' nearest neighbors among $S \cup T$ include fewer than k source examples. It is easy to see that this will always result in a (k, k')-NN-cover of T. Furthermore, this approach has a *query safety* property: the set T^l produced by Algorithm 2 satisfies $T^l \cap Q = \emptyset$ where $Q = \{x \in T : |k'(x, S \cup T) \cap S| \geq k\}$ is the set of target examples that have k source neighbors among their k' nearest neighbors in $S \cup T$. In other words, Algorithm 2 will not query the label of any target example in regions with sufficiently many labeled source examples nearby, a property used in the query guarantee of Theorem 2.

In order to make as few label queries as possible, we would like to find the smallest subset T^l of T to be labeled such that $T^l \cup S$ is a (k, k')-NN-cover of T. This problem is a special case of MINIMUM MULTISET MULTICOVER, a generalization of the well-known NP-hard MINIMUM SET COVER problem (see [4], Chapter 13.2). In Appendix B we discuss this problem further and give an efficient approximation algorithm we call EMMA.

While EMMA does not have the same query safety property enjoyed by Safe, we can ensure that an intelligent query strategy like EMMA still has the desired query safety property by first running Safe and then passing the resulting set T_{safe} to EMMA as its unlabeled sample. We call the resulting strategy for finding a (k, k')-NN-cover Safe-EMMA.

3 Performance guarantees

In this section, we analyze the expected loss of the output classifier of ANDA as well as its querying behavior. The bound in Section 3.1 on the loss holds for ANDA with any of the sub-procedures presented in Section 2. To simplify the presentation we use ANDA as a placeholder for any of ANDA-Safe , ANDA-EMMA and ANDA-Safe-EMMA . The bounds on the number of queries in Section 3.2 hold for ANDA-Safe and ANDA-Safe-EMMA , which we group under the placeholder ANDA-S . The proofs of all results in this section have been moved to the appendix.

3.1 Bounding the loss

We now provide bounds on the loss of the output classifier of Algorithm 1. We start with presenting finite sample bounds under the assumption that the regression function η satisfies a λ-Lipschitz condition. That is, we have $|\eta(x) - \eta(x')| \leq \lambda \rho(x, x')$ for all $x, x' \in \mathcal{X}_S \cup \mathcal{X}_T$.

Our bound on the expected loss in Theorem 1 is shown using standard techniques for nearest neighbor analysis. However, since our algorithm does not predict with a fully labeled sample from the target distribution (possibly very few of the target generated examples get actually labeled and the prediction is mainly based on source generated examples), we need to ensure that the set of labeled examples still sufficiently covers the target task. The following lemma serves this purpose. It bounds the distance of an arbitrary domain point x to its k-th nearest *labeled point* in terms of its distance to its k'-th nearest *target sample point*. Note that the bound in the lemma is easy to see for points in T. However, we need it for arbitrary (test-) points in the domain.

Lemma 1. *Let T be a finite set of points in a metric space $(\mathcal{X}, \rho)$ and let R be a (k, k')-*NN*-cover for T. Then, for all $x \in \mathcal{X}$ we have*

$$\rho(x, x_k(x, R)) \leq 3\rho(x, x_{k'}(x, T))$$

This lemma allows us to establish the finite sample guarantee on the expected loss of the classifier output by ANDA. Note that the guarantee in the theorem below is independent of the size and the generating process of S (except for the labels being generated according to η), while possibly (if S covers the target sufficiently) only few target points are queried for labels. Recall that $\mathrm{N}_\epsilon(\mathcal{X}, \rho)$ denotes the ϵ-covering number of a metric space.

Theorem 1. *Let $(\mathcal{X}, \rho)$ be a metric space and let P_T be a (target) distribution over $\mathcal{X} \times \{0, 1\}$ with λ-Lipschitz regression function η. Then for all $k' \geq k \geq 10$, all $\epsilon > 0$, and any unlabeled sample size m_T and labeled sequence $S = ((x_1, y_1), \ldots, (x_{m_S}, y_{m_S}))$ with labels y_i generated by η,*

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

The proof employs standard techniques (as in [5]) and incorporates our bound on the k nearest labeled points of Lemma 1. We also prove that ANDA is *consistent* in a slightly more general setting, namely if the regression function is *uniformly continuous* and the $\mathrm{N}_\epsilon(\mathcal{X}_T, \rho)$ are finite. This result can be found in Appendix C.

3.2 Bounding the number of queries

In this section, we show that our algorithm automatically adapts the number of label queries to the similarity of source and target task. We provide finite sample bounds that imply that with a sufficiently large source sample, with high probability, ANDA-S does not query at all in areas where the weight ratio of balls is bounded from below; i.e. it only queries where it is "needed." Recall that $B_{k,T}(x)$ denotes the smallest ball around x that contains the k nearest neighbors of x in T and $\beta(B) = D_S(B)/D_T(B)$ is the weight ratio. Let $\mathcal{B}$ denote the class of balls in $(\mathcal{X}, \rho)$.

Theorem 2. *Let $\delta > 0$, $w > 0$ and $C > 1$. Let m_T be some target sample size and let the source sample size satisfy*

$$m_S \geq \max\left\{4\left(\frac{3\,\mathrm{VCdim}(\mathcal{B})\,m_T}{C\,k\,w}\right)\ln\left(\frac{3\,\mathrm{VCdim}(\mathcal{B})\,m_T}{C\,k\,w}\right),\ \frac{3\,\ln(6/\delta)\,m_T}{C\,k\,w},\ \frac{9\,m_T}{C\,w}\right\},$$

for some k that satisfies $k \geq 9\,(\mathrm{VCdim}(\mathcal{B})\ln(2m_T) + \ln(6/\delta))$ and $m_T > k' = (C+1)k$. Then, with probability at least $1 - 2\delta$ over samples S of size m_S (i.i.d. from P_S) and T of size m_T (i.i.d. from D_T), ANDA-S *on input S, T, k, k' will not query any points $x \in T$ with $\beta(B_{Ck,T}(x)) > w$.*

Theorem 2 implies that if the source and target distributions happen to be identical or very close then, given that ANDA-S is provided with a sufficiently large source sample, it will not make any label queries at all. The same holds true if the weight ratio between source and target is uniformly lower bounded by some constant $w > 0$. Moreover, the theorem shows that, independent of an overall source/target relatedness measure, the querying of ANDA-S adapts automatically to a *local* relatedness measure in form of a weight ratio of balls around target sample points. ANDA-S queries only in areas that are not sufficiently supported by the source, that is, in areas where it is needed.

4 Experiments

Our experiments on synthetic data illustrate ANDA's adaptation ability and show that its classification performance compares favorably with baseline passive nearest neighbor methods. See Appendix E for experimental methods and results.

References

[1] R. Chattopadhyay, Z. Wang, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Batch mode active sampling based on marginal probability distribution matching," *TKDD*, vol. 7, no. 3, p. 13, 2013.

[2] R. Chattopadhyay, W. Fan, I. Davidson, S. Panchanathan, and J. Ye, "Joint transfer and batch-mode active learning," in *ICML*, pp. 253–261, 2013.

[3] A. Saha, P. Rai, H. D. III, S. Venkatasubramanian, and S. L. DuVall, "Active supervised domain adaptation," in *ECML/PKDD*, pp. 97–112, 2011.

[4] V. Vazirani, *Approximation Algorithms*. Springer, 2001.

[5] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*. Cambridge University Press, 2014.

[6] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, pp. 1345–1359, October 2010.

[7] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, pp. 55–66, 2010.

[8] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of representations for domain adaptation," in *NIPS*, pp. 137–144, 2006.

[9] Y. Mansour, M. Mohri, and A. Rostamizadeh, "Domain adaptation: Learning bounds and algorithms," in *Proceedings of the Conference on Learning Theory (COLT)*, 2009.

[10] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariate shift adaptation," *Annals of the Institute of Statistical Mathematics*, vol. 60, no. 4, pp. 699–746, 2008.

[11] S. Ben-David and R. Urner, "Domain adaptation-can quantity compensate for quality?," *Ann. Math. Artif. Intell.*, vol. 70, no. 3, pp. 185–202, 2014.

[12] C. Cortes, Y. Mansour, and M. Mohri, "Learning bounds for importance weighting," in *Advances in Neural Information Processing Systems (NIPS)* (J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, eds.), pp. 442–450, 2010.

[13] Y. Shi and F. Sha, "Information-theoretical learning of discriminative clusters for unsupervised domain adaptation," in *ICML*, 2012.

[14] S. Dasgupta, "Two faces of active learning," *Theor. Comput. Sci.*, vol. 412, no. 19, pp. 1767–1781, 2011.

[15] S. Dasgupta, "Analysis of a greedy active learning strategy," in *Advances in Neural Information Processing Systems (NIPS)*, pp. 337–344, MIT Press, 2004.

[16] M.-F. Balcan, A. Broder, and T. Zhang, "Margin based active learning," *Proceedings of the Conference on Learning Theory (COLT)*, pp. 35–50, 2007.

[17] M.-F. Balcan, A. Beygelzimer, and J. Langford, "Agnostic active learning," *J. Comput. Syst. Sci.*, vol. 75, no. 1, 2009.

[18] S. Hanneke, "Rates of convergence in active learning," *The Annals of Statistics*, vol. 39, no. 1, pp. 333–361, 2011.

[19] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[20] C. J. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, vol. 5, pp. 595–620, 07 1977.

[21] S. R. Kulkarni and S. E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," *IEEE Transactions on Information Theory*, vol. 41, no. 4, pp. 1028–1039, 1995.

[22] S. Kpotufe, "k-nn regression adapts to local intrinsic dimension," in *NIPS*, pp. 729–737, 2011.

[23] S. Dasgupta and K. Sinha, "Randomized partition trees for exact nearest neighbor search," in *COLT*, pp. 317–337, 2013.

[24] P. Ram, D. Lee, and A. G. Gray, "Nearest-neighbor search on a time budget via max-margin trees," in *SDM*, pp. 1011–1022, 2012.

[25] S. Dasgupta, "Consistency of nearest neighbor classification under selective sampling," in *COLT*, pp. 18.1–18.15, 2012.

[26] S. Rajagopalan and V. V. Vazirani, "Primal-dual RNC approximation algorithms for (multi)-set (multi)-cover and covering integer programs," in *Proceedings of IEEE Symposium on Foundations of Computer Science*, pp. 322–331, 1993.

[27] Y. Zhao and S.-H. Teng, "Combinatorial and spectral aspects of nearest neighbor graphs in doubling dimensional and nearly-euclidean spaces," in *TAMC*, pp. 554–565, 2007.

[28] V. N. Vapnik and A. J. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability & Its Applications*, vol. 16, no. 2, pp. 264–280, 1971.

Acknowledgments

This work was supported in part by NSF grants CCF-1451177, CCF-1101283, and CCF-1422910, ONR grant N00014-09-1-0751, and AFOSR grant FA9550-09-1-0538.

A Related work

There is a rich body of applied studies for transfer or domain adaptation learning [6], and on selective sampling or active learning [7]. As mentioned in the introduction, there are also some recent studies that incorporate active learning strategies to deal with change in distributions. We thus focus our discussion on studies that formally analyze such algorithms and provide performance guarantees now.

For domain adaptation, even under covariate shift, performance guarantees usually involve an extra additive term that measures the difference between source and target tasks (that is the loss does not converge to the target optimal opt_T but to $opt_T + \Delta$, where Δ is some measure of distance between distributions) [8, 9], or they rely on strong assumptions, such as the target support being a subset of the source support and the density ratio between source and target being bounded from below [10, 11]. The case where the target is partly supported in regions that are not supported by the source (that is, there is no bound on the density ratio), is considered to be more realistic [12], yet also particularly challenging. There are heuristics, that aim to find a suitable mapping of source and target into some common space [13], but the success of any such method again relies on very strong prior knowledge about source and target relatedness. We show that our method guarantees small loss independently of any source target relatedness.

The theory of active learning has received a lot of attention in recent years (see [14] for a survey on the main directions). Active learning has mostly been studied in a parametric setting (that is, learning some hypothesis class H) and benefits and limitations of active querying strategies have been proven in the realizable setting [15, 16] as well as in the agnostic case [17, 18]. However, the main goal incorporating active queries in all these works is to learn a classifier with low error while using fewer labels. In contrast, we focus on a different aspect of potential benefits of incorporating active queries and formally establish that being active is also useful to adapt to changing environments.

Nearest neighbor methods have been studied for decades [19, 20, 21]. Due to their flexibility, nearest neighbor methods suffer from a curse of dimension, both computationally and statistically. However, recently, there has been renewed interest in these methods and ways to overcome the curse of dimensionality. It has been proven that the generalization performance actually scales with notions of intrinsic dimension, which can be lower than the dimension of the feature space [22]. Several recent studies have shown how to perform nearest neighbor search more efficiently [23, 24]. Selective sampling for nearest neighbor classification has been shown to be consistent under certain conditions on the querying rule [25]; however, this work considers a data stream that comes from a fixed distribution (as opposed to our covariate shift setting). A 1-nearest neighbor algorithm has been analyzed under covariate shift [11]; however, that study assumes a fixed lower bound on a weight ratio between source and target, and therefore does not apply to settings where the target is supported in areas where the source is not. In our work, we argue that the flexibility of nearest neighbor methods (or non-parametric methods in general) can be utilized for being adaptive to changing environments; particularly so for choosing where to query for labels by detecting areas of the target task that are not well supported by the source.

B Finding a small (k, k')-NN-cover

In order to make as few label queries as possible, we would like to find the smallest subset T^l of T to be labeled such that $T^l \cup S$ is a (k, k')-NN-cover of T. This problem is a special case of MINIMUM MULTISET MULTICOVER, a generalization of the well-known NP-hard MINIMUM SET COVER problem (see [4], Chapter 13.2).

Definition (MINIMUM MULTISET MULTICOVER)**.** Given a universe U of n elements, a collection of multisets $\mathcal{S}$, and a coverage requirement r_e for each element $e \in U$, we say that a multiset $S \in \mathcal{S}$ covers element e once for each copy of e appearing in S. The goal is to find the minimum cardinality set $\mathcal{C} \subseteq \mathcal{S}$ such that every element $e \in U$ is covered at least r_e times by the multisets in $\mathcal{C}$.

Algorithm 3 EMMA: Efficient multiset multicover approximation for finding a (k, k')-NN-cover

input Labeled sample S, unlabeled sample T, parameters k, k'
$T^l \leftarrow \emptyset$
for all $x \in T$ **do**
  $r_x \leftarrow \max(0, k - k'(x, T \cup S) \cap S)$
  $n_x \leftarrow |\{x' \in T : r_{x'} > 0 \wedge x \in k'(x', S \cup T)\}|$
end for
while $\{x \in T : r_x > 0\} \neq \emptyset$ **do**
  $T^l \leftarrow T^l \cup \operatorname{argmax}_{x \in T \setminus T^l} r_x + n_x$
  for all $x \in T$ **do**
    $r_x \leftarrow \max(0, k - k'(x, T \cup S) \cap (S \cup T^l))$
    $n_x \leftarrow |\{x' \in T \setminus T^l : r_{x'} > 0 \wedge x \in k'(x', S \cup T)\}|$
  end for
end while
return T^l

We can phrase the problem of finding the smallest T^l such that $T^l \cup S$ is a (k, k')-NN-cover of T as a MINIMUM MULTISET MULTICOVER problem as follows. Let $U = T$ and set the coverage requirements as $r_x = \max(0, k - |k'(x, S \cup T) \cap S|)$ for each $x \in T$. The collection $\mathcal{S}$ contains a multiset S_x for each $x \in T$, where S_x contains k copies of x and one copy of each element in $\{x' \in T : x \in k'(x', S \cup T)\}$. By construction, a minimum multiset multicover of this instance is also a minimum (k, k')-NN-cover and vice versa.

While MINIMUM MULTISET MULTICOVER is NP-hard to solve exactly, a greedy algorithm efficiently provides an approximate solution (see Section B.1). Algorithm 3 formalizes this as an ANDA subroutine called EMMA for finding a small (k, k')-NN-cover. In the language of (k, k')-NN-covers, in each round EMMA computes the helpfulness of each $x \in T$ in two parts. The remaining coverage requirement r_x is the number of times x would cover itself if added to T^l (that is, the savings from not having to use r_x additional neighbors of x), and the total neighbor coverage n_x is the number of times x would cover its neighbors if added to T^l. EMMA then selects the point x with the largest sum $r_x + n_x$ among all points in T that have not yet been added to T^l.

In its most basic form, EMMA does not have the same query safety property enjoyed by Safe because the greedy strategy may elect to query labels of target examples that were already fully covered by source examples. We can ensure that an intelligent query strategy like EMMA still has the desired query safety property by first running Safe and then passing the resulting set T_{safe} to EMMA as its unlabeled sample. We call the resulting strategy for finding a (k, k')-NN-cover Safe-EMMA.

B.1 Approximation guarantees

MINIMUM MULTISET MULTICOVER is known to remain NP-hard even when the multisets in $\mathcal{S}$ are small. However, a small upper bound b on the maximum size of any multiset in $\mathcal{S}$ can make the problem much easier to approximate. Specifically, the greedy algorithm has an approximation factor of H_b, the b-th harmonic number [26]. This is known to be essentially optimal under standard hardness assumptions.

In our setting, the size of the largest multiset is determined by the point $x \in T$ with the largest number of points in $S \cup T$ having x as one of their k' nearest neighbors. In general metric spaces this can be up to $m = m_S + m_T$, resulting in a multiset of size $m + k$ and an approximation factor of $H_{m+k} = O(\log m)$. However, in spaces with doubling-dimension γ, it is known that $b \leq k' 4^\gamma \log_{3/2}(2L/S)$ where L and S are respectively the longest and shortest distances between any two points in T [27].

C Consistency

We show that ANDA is consistent in a slightly more general setting, namely if the regression function is *uniformly continuous* and the $\mathrm{N}_\epsilon(\mathcal{X}, \rho)$ are finite. Note that this is the case, for example, if

$(\mathcal{X}, \rho)$ is compact and η is continuous. Recall that, a function $\eta : \mathcal{X} \to \mathbb{R}$ is *uniformly continuous* if for every $\gamma > 0$ there exists a δ such that for all $x, x' \in \mathcal{X}$, $\rho(x, x') \leq \delta \Rightarrow |\eta(x) - \eta(x')| \leq \gamma$.

Corollary 1. *Let $(\mathcal{X}, \rho)$ be a metric space with finite covering numbers $\mathrm{N}_\epsilon(\mathcal{X}, \rho)$, and let $\mathcal{P}(\mathcal{X}, \rho)$ denote the class of distributions over $\mathcal{X} \times \{0, 1\}$ with uniformly continuous regression functions. Let $(k_i)_{i\in\mathbb{N}}$, $(k'_i)_{i\in\mathbb{N}}$ and $(m_i)_{i\in\mathbb{N}}$ be non-decreasing sequences of natural numbers with $k'_i \geq k_i$ for all i, and $k_i \to \infty, k'_i \to \infty, m_i \to \infty$ and $(k'_i/m_i) \to 0$ as $i \to \infty$. Let $(S_i)_{i\in\mathbb{N}}$ be a sequence of labeled domain points, that is for all i we have $S_i \in (\mathcal{X} \times \{0, 1\})^n$ for some n. Then for any distribution $P_T \in \mathcal{P}(\mathcal{X}, \rho)$, we have*

$$\lim_{i\to\infty} \mathop{\mathbb{E}}_{T\sim P_T^{m_i}} [\mathcal{L}_T(\mathrm{ANDA}(S_i, T, k_i, k'_i))] \;=\; \mathcal{L}_T(h^*).$$

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

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

for all $k \geq 10$ and $k' \geq k$.

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. □

D Proofs

D.1 Proof of Lemma 1

Proof. Let $x \in \mathcal{X}$. If the set $k'(x, T)$ of the k' nearest neighbors of x in T contains k points from R, we are done (in this case we actually have $\rho(x, x_k(x, R)) \leq \rho(x, x_{k'}(x, T))$). Otherwise, let $x' \in k'(x, T) \setminus R$ be one of these points that is not in R. Since R is a (k, k')-NN-cover for T, and $x' \in T$, the set of the k' nearest neighbors of x' in $R \cup T$ contains k elements from R.

Let x'' be any of these k elements, that is $x'' \in R \cap k'(x', R \cup T)$. Note that $\rho(x', x'') \leq 2\rho(x, x_{k'}(x, T))$ since x' is among the k' nearest neighbors of x and x'' is among the k' nearest neighbors of x' in $R \cup T$. Thus, we have

$$\rho(x, x'') \leq \rho(x, x') + \rho(x', x'') \leq \rho(x, x_{k'}(x, T)) + 2\rho(x, x_{k'}(x, T)) = 3\rho(x, x_{k'}(x, T))$$

which completes the proof. □

D.2 Proof of Theorem 1

We adapt the proof (guided exercise) of Theorem 19.5 in [5] 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 2 (Lemma 19.6 in [5])**.** *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 3 (Lemma 19.7 in [5]). *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 [5]). *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:

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

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 2 implies that this first summand is bounded in expectation by $\frac{2\,\mathrm{N}_\epsilon(\mathcal{X},\rho)\,k'}{m_T}$.

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 $k(x, L) = \{x_1, \ldots, x_k\}$ 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

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

where the first inequality follows from Claim 1 and the second from Lemma 3. We have

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

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

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

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

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

where the rearrangement in the first two steps is by Fubini's theorem. 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. □

D.3 Proof of Theorem 2

In our analysis, we employ Lemma 4 below. It follows from VC-theory [28] and appears in [22]. We let $\widehat{S}, \widehat{T}$ denote empirical distributions according to source or target sample S and T, respectively.

Lemma 4 (Lemma 1 in [22])**.** *Let $\mathcal{B}$ denote the class of balls in $(\mathcal{X}, \rho)$, and let D be a distribution over $\mathcal{X}$. Let $0 < \delta < 1$, and define $\alpha_n = (\mathrm{VCdim}(\mathcal{B})\ln(2n) + \ln(6/\delta))/n$. The following holds with probability at least $1-\delta$ (over a sample T of size n drawn i.i.d. from D) for all balls $B \in \mathcal{B}$: if $a \geq \alpha_n$, then $\widehat{T}(B) \geq 3a$ implies $D(B) \geq a$ and $D(B) \geq 3a$ implies $\widehat{T}(B) \geq a$.*

With this we can prove the query bound of Theorem 2.

Proof. Note that $m_S \geq 4\left(\frac{3\,\mathrm{VCdim}(\mathcal{B})\,m_T}{C\,k\,w}\right)\ln\left(\frac{3\,\mathrm{VCdim}(\mathcal{B})\,m_T}{C\,k\,w}\right)$, implies that $m_S \geq \left(\frac{3\,\mathrm{VCdim}(\mathcal{B})\,m_T}{C\,k\,w}\right)\ln(2m_S)$, and together with the second lower bound (in the max) on m_S, this yields

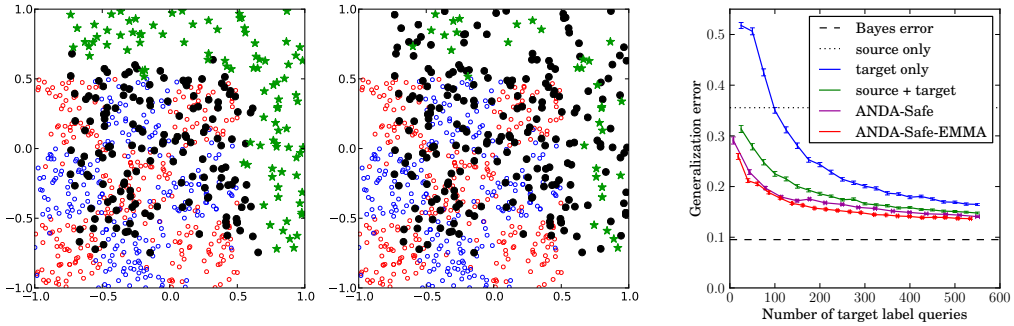
$$m_S\frac{C\,k\,w}{3\,m_T} \geq \mathrm{VCdim}(\mathcal{B})\ln(2m_S) + \ln(6/\delta). \tag{1}$$

We now assume that S and T are so that the implications in Lemma 4 are valid (this holds with probability at least $1-2\delta$ over the samples S and T). Let $x \in T$ be such that $\beta(B_{Ck,T}(x)) > w$. By definition of the ball $B_{Ck,T}(x)$, we have $\widehat{T}(B_{Ck,T}(x)) = \frac{Ck}{m_T}$, and by our choice of k, therefore

$$\widehat{T}(B_{Ck,T}(x)) = \frac{C\,k}{m_T} \geq \frac{C\,9\,(\mathrm{VCdim}(\mathcal{B})\ln 2m_T + \ln 6/\delta)}{m_T}.$$

Now Lemma 4 implies that $D_T(B_{Ck,T}(x)) \geq \frac{C\,k}{3\,m_T}$, so the condition on the weight ratio of this ball now yields

$$D_S(B_{Ck,T}(x)) \geq \frac{C\,k\,w}{3\,m_T} = m_S\frac{C\,k\,w}{3\,m_T\,m_S} \geq \frac{\mathrm{VCdim}(\mathcal{B})\ln(2m_S) + \ln(6/\delta)}{m_S},$$



(a) (b)

Figure 1: (a) Visualization of snythetic data and query strategies for ANDA-Safe (left) and ANDA-Safe-EMMA (right). Red and blue circles represent labeled source examples, black circles represent unqueried target examples, and green stars represent queried target examples. (b) Experimental results comparing our approach to baseline methods. Error bars represent two standard errors, or roughly a 95% confidence interval.

where the last inequality follows from Equation (1). Now, Lemma 4, together with the assumption $m_S \geq \frac{9\,m_T}{C\,w}$ (the third term in the max), implies $\widehat{S}(B_{Ck,T}(x)) \geq \frac{C\,k\,w}{9\,m_T} \geq \frac{k}{m_S}$. This means that $B_{Ck,T}(x)$ contains k examples from the source, which implies that among the $k' = Ck + k$ nearest sample points (in $S \cup T$) there are k source examples, and therefore x will not be queried by ANDA-S . □

E Experiments

Our experiments on synthetic data demonstrate ANDA's adaptation ability and show that its classification performance compares favorably with baseline passive nearest neighbor methods. The source marginal D_S was taken to be the uniform distribution over $[-1, 0.5]^2$ and the target marginal D_T was set to uniform over $[-0.75, 1]^2$. This ensures enough source/target overlap so the source data is helpful in learning the target task but not sufficient to learn well. The regression function chosen for both tasks was $\eta(x, y) = (1/2)(1 - (\sin(2\pi x)\sin(2\pi y))^{1/6})$ for $(x, y) \in \mathbb{R}^2$. This creates a 4×4 checkerboard of mostly-positively and mostly-negatively labeled regions with noise on the boundaries where η crosses $1/2$. Training samples from this setting are pictured in Figure 1a.

The baseline algorithms we compare against are the following. The "source only" algorithm predicts according to a k-NN classifier built on a source sample alone. The "target only" algorithm creates a k-NN classifier on a random sample from the target, and the "source + target" does the same but includes labeled data from a source sample as well.

We compare the generalization error of ANDA-Safe-EMMA and ANDA-Safe against these baselines across a range of unlabeled target sample sizes. Since the number of queries made by both ANDA-Safe-EMMA and ANDA-Safe increases with target sample size, this generates a range of query counts for the active algorithms. The baseline algorithms were given labeled target samples of sizes in the same range as these query counts. For all algorithms and target sample sizes we fixed $m_S = 3200$, $k = 7$ and $k' = 21$. Figure 1b shows the resulting generalization error for each algorithm as a function of the number of target labels used. Each point in the plot represents an average over 100 independent trials.

Both active algorithms perform significantly better than the passive baselines in terms of the error they achieve per target label query. ANDA-Safe-EMMA outperforms ANDA-Safe as well, since (as demonstrated in Figure 1a) it can achieve full coverage of the target region with many fewer queries.