
Margin-Independent Online Multiclass Learning via Convex Geometry

Guru Guruganesh
Google Research
gurug@google.com

Allen Liu
MIT
cliu568@mit.edu

Jon Schneider
Google Research
jschnei@google.com

Joshua Wang
Google Research
joshuawang@google.com

Abstract

We consider the problem of multi-class classification, where a stream of adversarially chosen queries arrive and must be assigned a label online. Unlike traditional bounds which seek to minimize the misclassification rate, we minimize the total distance from each query to the region corresponding to its correct label. When the true labels are determined via a nearest neighbor partition – i.e. the label of a point is given by which of k centers it is closest to in Euclidean distance – we show that one can achieve a loss that is independent of the total number of queries. We complement this result by showing that learning general convex sets requires an almost linear loss per query. Our results build off of regret guarantees for the geometric problem of contextual search. In addition, we develop a novel reduction technique from multiclass classification to binary classification which may be of independent interest.

1 Introduction

Online multiclass classification is a ubiquitous problem in machine learning. In this problem, a learning algorithm is presented with a stream of incoming query points and is tasked with assigning each query with a label from a fixed set. After choosing a label, the algorithm is told the true label of the query point. The goal of the algorithm is to learn over time how to label the query points as accurately as possible.

Traditionally, theoretical treatments of this problem are built around the notion of a *margin* γ . This margin represents the extent to which the input points are well-separated from the boundaries between different labels. For example, the analysis of the classic Perceptron algorithm [Nov63] guarantees that it makes at most $O(1/\gamma^2)$ mistakes when performing binary classification, as long as all query points are distance at least γ from a hyperplane separating the two classes. More sophisticated analyses and algorithms (relying on e.g. hinge loss) do not necessarily assume the classes are as well separated, but still inherently incorporate a margin γ (for example, the hinge loss associated with a point is positive unless it is γ -separated).

In this paper, we present an alternative to the traditional margin approaches. Our approach weights each mistake by how ambiguous the classification task is for that point, rather than penalizing all mistakes equally. More precisely, consider a partition of the space of all possible query points into k regions R_i , where R_i contains all query points whose true label is i . In our formulation assigning a query point q a label i incurs a loss of $\ell(q, R_i)$, where $\ell(q, R_i)$ should be thought of as *the distance needed to move q so that it lies in R_i* (i.e., for it to be labelled correctly). For example, in the case of

a linear classifier, $\ell(q, R_i)$ is zero if q is correctly classified, and the distance to the classifier if q is incorrectly classified. The goal of the algorithm is to minimize the total loss.

This notion of loss not only measures the rate of errors but also the degree of each error; choosing a wildly inaccurate label is punished more than selecting a label that is "almost" correct. This fine-grained approach to looking at errors has occurred in other areas of machine learning research as well. For example, the technique of knowledge distillation is based on training a smaller model on the logits produced by a larger model [HVD15]. Hinton et al. explain, "The relative probabilities of incorrect answers tell us a lot about how the cumbersome model tends to generalize. An image of a BMW, for example, may only have a very small chance of being mistaken for a garbage truck, but that mistake is still many times more probable than mistaking it for a carrot." Rather than leaning on the power of a trained model, our framework differentiates between these different incorrect answers based on the geometry of the problem.

1.1 Our Results

1.1.1 Learning Linear Classifiers

In this case we have a binary classification problem, where the two regions R_1 and R_2 are separated by an unknown d -dimensional hyperplane $\langle v, x \rangle = 0$ (with $\|v\|_2 = 1$). Our loss function in this case is the function $\ell(q, R_i) = |\langle q, v \rangle| \cdot \mathbf{1}(q \notin R_i)$. We prove the following result:

Theorem 1 (Restatement of Corollary 3.1). *There exists an efficient algorithm for learning a linear classifier that incurs a total loss of at most $O(d \log d)$.*

Note that the total loss in Theorem 1 is *independent* of the time horizon (number of rounds) T . More importantly, note that this is stronger than the naive guarantee implied by the margin bounds for the Perceptron algorithm. Indeed, each mistake at a distance γ from the separating hyperplane is assigned loss $O(\gamma)$ in our model. Since the Perceptron algorithm can make up to $O(1/\gamma^2)$ such mistakes, this only implies Perceptron incurs a total loss of at most $O(1/\gamma)$ (which blows up as $\gamma \rightarrow 0$).

Indeed, our algorithm in Theorem 1 is not based off of the Perceptron algorithm or its relatives, but rather off of recently developed algorithms for a problem in online learning known as *contextual search*. In contextual search, a learner is similarly faced with a stream of incoming query points q_t and wishes to learn a hidden vector v . However, instead of trying to predict the sign of $\langle v, q_t \rangle$, in contextual search the goal is to guess the *value* of $\langle v, q_t \rangle$. After the learner submits a guess, they are told whether or not their guess was higher or lower than the true value of $\langle v, q_t \rangle$ (and pay a loss equal to the distance between their guess and the truth). The best known contextual search algorithms rely on techniques from integral geometry (bounding various intrinsic volumes of the allowable knowledge set), and are inherently different than existing Perceptron/SVM-style algorithms.

While it may seem like contextual search (which must predict the value of $\langle v, q_t \rangle$ instead of just the sign) is strictly harder than our online binary classification problem, they are somewhat incomparable (for example, unlike in contextual search, we have no control over what feedback we learn about the hidden vector v). Nonetheless, in Theorem 8 we show a general reduction from our binary classification problem to contextual search. This allows us to use recent results of [LLS20] to obtain our $O(d \log d)$ bound in Theorem 1.

1.1.2 Learning Nearest Neighbor Partitions

One natural way to split a query space into multiple classes is via a *nearest neighbor partition*. In this setting, each label class i is associated with a "center" $x_i \in \mathbb{R}^d$, and each region R_i consists of the points which are "nearest" to x_i . To define "nearest", we introduce a similarity metric $\delta(x, y)$ representing the "distance" between points x and y in \mathbb{R}^d . The two major classes of similarity metrics we consider are: a) the *inner-product similarity* $\delta(x, y) = -\langle x, y \rangle$ and b) the *L^p similarity* $\delta(x, y) = \|x - y\|_p$. Given a fixed similarity metric δ , our loss function in this case is the function $\ell(q, R_i) = \delta(q, x_i) - \min_{i^*} \delta(q, x_{i^*})$; in other words, the difference between the similarity between q and x_i with the similarity between q and its most similar center.

Theorem 2 (Restatement of Corollary 3.3). *For inner-product similarity, there exists an efficient randomized algorithm for learning a nearest neighbors partition that incurs a total expected loss of at most $O(k^2 d \log d)$.*

Like some other algorithms for multiclass classification, our algorithm in Theorem 2 works by running one instance of our binary classification algorithm (Theorem 1) for each pair of labels. Unlike some other “all-vs-all” methods in the multiclass classification literature, however, it does not suffice to run a simple majority vote over these instances. Instead, to prove Theorem 2, we solve a linear program to construct a probability distribution over centers that guarantees that our expected loss is bounded by an expected decrease in a total potential of all our $\binom{k}{2}$ sub-algorithms.

Our results for L^p similarity are as follows:

Theorem 3 (Restatement of Theorems 12 and 13). *For L^p similarity, when p is a positive even integer, there exists an efficient randomized algorithm for learning a nearest neighbors partition that incurs a total expected loss of at most $O(k^2 \text{poly}(p, d))$.*

For an arbitrary $p \geq 2$, if all k centers are Δ -separated in L^p distance, there exists an efficient randomized algorithm that incurs a total expected loss of

$$\frac{k^2 \text{poly}(p, d)}{\Delta} \cdot \left(\frac{1}{p-2} \right)^2.$$

When p is an even integer, it is possible to construct a polynomial kernel that exactly reduces this problem to the problem for inner-product similarity (albeit in the higher dimensional space $\mathbb{R}^{d(p+1)}$). When p is not an even integer, it is no longer possible to perform an exact reduction to the inner-product similarity problem. Instead, in parallel we perform a series of approximate reductions to inner-product similarity at multiple different scales (the full algorithm can be found in Appendix¹ E). Surprisingly, this technique only gives T -independent bounds on the loss when $p \geq 2$. It is an interesting open problem to develop algorithms for the case $1 \leq p < 2$ (and more generally, for arbitrary norms).

1.1.3 Learning General Convex Regions

Finally, we consider the case where the regions R_i are not defined in relation to hidden centers, but where they can be any convex subsets of \mathbb{R}^d . Interestingly, in this case it is impossible to achieve total loss independent of T . Indeed, we prove a lower bound of $\Omega(T^{1-O(1/d)})$ for the total loss of any algorithm for this setting, even when $k = 2$.

Theorem 4. *Any algorithm for learning general convex regions incurs a total loss of at least $\Omega(T^{(d-4)/(d-2)})$, even when there are only two regions.*

1.2 Mistake Bounds and Halving Algorithms

As mentioned in the introduction, classical algorithms for multi-class classification generally try to minimize the *mistake bound* (the total number of classification errors the algorithm makes) under some margin guarantee γ . Even though our algorithms are designed to minimize the absolute loss and not a margin-dependent mistake bound, it is natural to ask whether our algorithms come with any natural mistake bound guarantees.

We show that our algorithms do in fact possess strong mistake bounds, matching the dependence on the margin γ of the best known halving algorithms. In particular, we show the following.

Theorem 5 (Restatement of Theorem 16). *If all query points q_t are at least distance γ away from the separating hyperplane, our algorithm for learning linear classifiers (Theorem 1) makes at most $O(d \log(d/\gamma))$ mistakes.*

Theorem 6 (Restatement of Theorem 17). *If all query points q_t are at least distance γ away from the boundary between any two regions, our algorithm for learning nearest neighbor partitions (Theorem 2) makes at most $O(k^2 d \log(d/\gamma))$ mistakes.*

In comparison, the best dimension-dependent classical bounds for this problem come from halving algorithms (efficiently implementable via linear programming) which have mistake bounds of $O(d \log(1/\gamma))$ and $O(kd \log(1/\gamma))$ respectively. In the first case, our mistake bound is nearly tight (losing only an additive $O(d \log d)$). In the second case, our mistake bound is tight up to a multiplicative factor of k ; it is an interesting open question whether it is possible to remove this factor of k in our techniques.

¹All references to the appendix refer to the appendices of the Supplementary Material.

We additionally introduce a variant on the mistake bound that we call a *robust mistake bound* that is defined as follows. Normally, when we have a margin constraint of γ , we insist that all query points q_t are distance at least γ away from any separating hyperplane. In our robust model, we remove this constraint, but only count mistakes when the query point q_t lies at least γ away from the separating hyperplane.

Existing algorithms (both the Perceptron and halving algorithms) do not appear to give any non-trivial mistake bounds in the robust model – it is very important for the analysis of these algorithms that *every* query point is far from the separating hyperplane. On the other hand, it straightforwardly follows from our notion of loss that if we have an $O(R)$ -loss algorithm for some problem, that algorithm simultaneously achieves an $O(R/\gamma)$ robust mistake bound. In particular, we obtain robust mistake bounds of $O((d \log d)/\gamma)$ and $O((k^2 d \log d)/\gamma)$ for learning linear classifiers and learning nearest neighbor partitions respectively.

Related work The problem of online binary classification (and specifically, of online learning of a linear classifier) is one of the oldest problems in machine learning. The Perceptron algorithm was invented in [Ros58], and the first mistake bound analysis of the Perceptron algorithm appeared in [Nov63]. Since then, there has been a myriad of research on this problem, some of which is well-surveyed in [MR13]. Of note, the first mistake bounds for the non-separable case appear in [FS99]. We are not aware of any work on this problem that investigates the same loss we present in this paper.

Bounds for support vector machines (see [Vap13]) also result in the use of a margin to bound the number of mistakes. Choosing a suitable kernel can help create or improve the margin when viewing the data in the “kernel” space. We use a similar technique in proving bounds for generalized L^p norms. Our idea is to use differently-scaled kernels to produce increasingly accurate approximations. To the best of our knowledge, the technique we present is novel. There are other related techniques in the literature, e.g. [SBD06] attempt to learn the best kernel to improve classification error.

Similarly, there is a wealth of both theoretical and empirical research on multiclass classification. As far back as 1973, researchers were looking at generalizing binary classification algorithms to this setting (see e.g. Kesler’s construction in [DH⁺73]). One piece of relevant work is [CS03], which generalizes online binary classification Perceptron-style algorithms to solve online multiclass classification problems – in particular, the multiclass hypotheses they consider are same as the nearest neighbor partitions generated by the inner-product similarity metric (although as with the Perceptron, they only analyze the raw classification error). Since then, this work has been extended in many ways to a variety of settings (e.g. [CDK⁺06, CG13, KSST08]).

Another way of looking at the problem of multiclass classification is that we are learning how to cluster a sequence of input points into k pre-existing clusters. Indeed, the nearest neighbor partition with L^2 similarity metric gives rise to exactly the same partitioning as a k -means clustering. There is an active area of research on learning how to cluster in an online fashion (e.g. [GLZ14, GLK⁺17, BR20]). Perhaps most relevantly, [LSS16] studies a setting where one must choose cluster labels for incoming points in an online fashion, and then at the end of the algorithm can choose the k centers (the goal being to minimize the total k -means loss of this eventual clustering).

The algorithms we develop in this paper are based off of algorithms for contextual search. Contextual search is a problem that originally arose in the online pricing community [CLPL16, LLV17]. The variant of contextual search we reference in this paper (with symmetric, absolute value loss) first appeared in [LS18]. The algorithms in this paper were later improved in [LLS20] (and it is these improved algorithms that we build off of in this paper).

2 Model

Notation We will write $B_d(c, r)$ to denote the d -dimensional ball in \mathbb{R}^d centered at c with radius r . We write B_d in place of $B_d(0, 1)$ to indicate the unit ball centered at the origin.

Given an $x \in \mathbb{R}^d$, we will write $\|x\|_p$ to denote the L^p norm of the vector x . In the case of $p = 2$, we will often omit the subscript and write $\|x\|$ in place of $\|x\|_2$.

2.1 Online Multiclass Learning

We will view the problem of online multiclass learning as follows. There are k disjoint regions in some domain, say B_d , labelled R_1 through R_k . The region R_i contains the points in B_d that should be assigned the label i . The goal of the learner is to learn these subsets (and thus how to label points in B_d in an online manner). Every round t , the learner receives an adversarially chosen query point $q_t \in B_d$. The learner must submit a prediction $I_t \in [k]$ for which region R_{I_t} the point q_t lies in. The learner then learns which region $R_{I_t^*}$ the point actually belongs to, and suffers some loss $\ell(q_t, R_{I_t})$. This loss function should in some way represent how far q_t was from lying in the region R_{I_t} chosen by the learner; for example, in the case where the learner chooses the correct region $R_{I_t^*}$, $\ell(q_t, R_{I_t^*})$ should be zero.

In this paper, we will consider two specific cases of the above learning problem. In the first case, we wish to learn a *nearest-neighbor partition*. That is, the k regions are defined by k ‘‘centers’’ $x_1, x_2, \dots, x_k \in B_d$. Region R_i then consists of all the points which are ‘‘nearest’’ to center x_i according to some similarity metric $\delta(x, y)$ (where lower values of $\delta(x, y)$ mean that x and y are more similar; note that $\delta(x, y)$ *does not* need to be an actual metric obeying the triangle-inequality). Given a similarity metric $\delta(x, y)$, the loss our algorithm incurs when labelling query q with label i is given by $\ell(q, R_i) = \delta(q, x_i) - \delta(q, x_{i^*})$, where R_{i^*} is the region containing query q .

We will examine several different possibilities for $\delta(x, y)$, including:

- $\delta(x, y) = -\langle x, y \rangle$. We refer to this as the *inner-product similarity* between x and y . Note that when $k = 2$, using this similarity metric reduces to the problem of learning a linear classifier. For $k > 2$, this results in similar partitions to those learned by multiclass perceptrons / SVMs [CS03, CDK⁺06].
- $\delta(x, y) = \|x - y\|_2$; in other words, the Euclidean distance between x and y . When using this loss function, the k regions are given by the Voronoi diagram formed by the k centers x_i .
- For $p \geq 1$, $\delta(x, y) = \|x - y\|_p$; in other words, the L^p distance between x and y .

There is a straightforward reduction from the Euclidean distance similarity to the inner-product similarity (see Appendix C), so in Section 3 we will primarily concern ourselves with the inner-product similarity. In Section 4 we will tackle this problem for the case of general L^p norms; for some cases (even integer p) it is possible to perform a similar reduction to the inner product similarity, but in general it is not and we will need to rely on other techniques.

In the second case, we wish to learn *general convex sets*. In particular, there are k disjoint convex sets $R_1, \dots, R_k \subseteq B_d$ (not necessarily a partition). Each round t , we receive a query point $q_t \in \bigcup_{i=1}^k R_i$, guess a region $i \in [k]$, and are penalized against the loss function $\ell(q_t, R_i) \triangleq \min_{x \in R_i} \|q_t - x\|_2$. In other words, we are penalized the minimum distance from q_t to the predicted region R_i , which is zero if our guess was correct. In Section 5 we will show that (even in the case of $k = 2$), there is no low-loss learning algorithm for learning general convex sets (in contrast to learning nearest neighbors).

Finally, for any $\alpha > 0$ and existing loss function $\ell(q, R_i)$, we can consider the modified loss function $\ell'(q, R_i) = \ell(q, R_i)^\alpha$. Note that this does not change the division of B_d into regions, but it *can* change the total loss incurred by our algorithm (and will be useful for some of our reductions). If in any result we do not specify an α , that means we are taking $\alpha = 1$ (i.e. the unmodified loss function).

2.2 Contextual Search

One of the main tools we will rely on is an existing algorithm for a problem in online learning known as *contextual search*. For our purposes, the problem of contextual search can be defined as follows. There is a hidden point $p \in B_d$, unknown to the learner. Every round t (for T rounds) an adversary provides the learner with a query vector $x_t \in B_d$. The learner must then submit a guess g_t for the value of the inner product $\langle x_t, p \rangle$. The learner then learns whether their guess was too high or too low. At the end of the game, the learner incurs loss $\ell(g_t, \langle x_t, p \rangle) = |g_t - \langle x_t, p \rangle|$ for each of their guesses. The learner’s goal is to minimize their total loss.

Interestingly, there exist algorithms for contextual search with total loss polynomial in the ambient dimension d and independent of the time horizon T . The first such algorithms were given in [LS18]

and used techniques from integral geometry to obtain a total regret of $O(d^4)$. More recently, these algorithms were improved in [LLS20] to achieve a regret bound of $O(d \log d)$. We will rely on a slightly strengthened variant of the result from [LLS20] to work when the loss function raised to an arbitrary power.

Theorem 7. *Let $\alpha > 0$. There exists an algorithm for contextual search with loss function $\ell(g_t, \langle x_t, p \rangle) = |g_t - \langle x_t, p \rangle|^\alpha$ that incurs a total loss of at most $O(\alpha^{-2} d \log d)$.*

In particular, Theorem 7 is satisfied by the algorithm from [LLS20]. For completeness, we include a description of the algorithm (along with the proof of Theorem 7) in Appendix A.

3 Learning Nearest Neighbor Partitions

3.1 The Two-Point Case: Learning a Hyperplane

To begin, we will discuss how to solve the $k = 2$ variant of the problem of learning nearest neighbor partitions for the inner-product similarity function $\delta(q, x) = -\langle q, x \rangle$. Recall that in this setting we have two unknown centers x_1 and x_2 belonging to B_d . Each round t we are given a query point $q_t \in B_d$, and asked to choose a label $I_t \in \{1, 2\}$ of the center that we think is most similar to q_t (i.e., that maximizes $\delta(q_t, x)$). Letting $y_t = x_{I_t}$ and $x_t^* = \arg \max_{x_i} \delta(q, x_i)$, our loss in round t is zero if we guess correctly ($y_t = x_t^*$) and is $\delta(q, x_t^*) - \delta(q, y_t)$ if we guess incorrectly (we will also be able to deal with the case when the loss is $|\delta(q, x_t^*) - \delta(q, y_t)|^\alpha$ for some $\alpha > 0$). Afterwards, we are told the identity (but not the location) of x_t^* .

Note that the optimal strategy in this game (for an agent who knows the hidden centers x_1 and x_2) is to guess $I_t = 1$ whenever $\langle q, x_1 \rangle > \langle q, x_2 \rangle$, and to guess $I_t = 2$ otherwise. Rewriting this, we want to guess $I_t = 1$ exactly when $\langle q, x_1 - x_2 \rangle > 0$. If we let $w = x_1 - x_2$, we can think of goal as learning the hyperplane $\langle q, w \rangle = 0$. More specifically, each round we are given a point q and asked which side of the hyperplane q lies on. If we guess correctly, we suffer no loss; if we guess incorrectly, we suffer loss equal to the distance from q to this hyperplane. In either case, we learn afterwards which side of the hyperplane q lies on.

3.1.1 Reducing to Contextual Search

We will show that we can solve this online learning problem with total loss $O(\text{poly}(d))$, independent of the number of rounds T in the time-horizon. As mentioned earlier, our primary tool will be existing algorithms for a problem in online learning known as *contextual search*.

Recall that in contextual search there is a hidden vector $v \in B_d$, and each round we are given a vector $q_t \in B_d$. However, unlike in our problem (where we only care about the sign of the inner product $\langle q_t, v \rangle$), the goal in contextual search is to submit a guess g_t for the value of the inner product $\langle q_t, v \rangle$. We then incur loss equal to the absolute distance $|\langle q_t, v \rangle - g_t|$ between our guess and the truth, and are then told whether our guess g_t was too high or too low.

As mentioned earlier in Section 2.2, there exist algorithms for contextual search with $O(d \log d)$ total loss. Via a simple reduction, we will show that we can apply these algorithms in our setting.

Theorem 8. *Fix $\alpha > 0$. Assume there exists an algorithm \mathcal{A} for contextual search with loss function $\ell(g_t, \langle x_t, p \rangle) = |g_t - \langle x_t, p \rangle|^\alpha$ that incurs regret at most $R(d, T)$. Then there exists an algorithm \mathcal{A}' that incurs regret at most $R(d, T)$ for the $k = 2$ case of learning nearest neighbor partitions with similarity metric $\delta(x, y) = -\langle x, y \rangle$ and loss raised to the power α .*

Corollary 3.1. *Fix an $\alpha > 0$. When $k = 2$, there exists an algorithm for learning nearest neighbor partitions with similarity metric $\delta(x, y) = -\langle x, y \rangle$ and loss raised to the power α that incurs total loss at most $O(\alpha^{-2} d \log d)$.*

3.1.2 Potential-Based Algorithms

In order to generalize this to $k > 2$ labels, we will need to open the black box that is our contextual search algorithm slightly. In particular, the argument in the following section requires our algorithm for the $k = 2$ case to be a *potential-based algorithm*.

Before defining exactly what a potential-based algorithm is, it will be useful to define the notion of a *knowledge set*. For the problem we are considering in this section – the $k = 2$ variant of our problem

for inner-product similarity – we will define the *knowledge set* K_t at time t to be the set of possible values for $w = x_1 - x_2$ that are consistent with all known information thus far. Note that since x_1 and x_2 start as arbitrary points in B_d , the knowledge set K_0 is simply the set $B_d - B_d = 2B_d$. As the algorithm gets more feedback about w , the knowledge set shrinks; however, since this feedback is always of the form of a linear constraint (e.g. $\langle q_t, w \rangle \geq 0$), the knowledge set K_t is always a convex subset of \mathbb{R}^d .

Let S_t be the history of all feedback the algorithm has seen up to (but not including) round t ; that is $S_t = \{(q_1, I_1^*), (q_2, I_2^*), \dots, (q_{t-1}, I_{t-1}^*)\}$. Let \mathcal{S}_t denote the set of possible values for S_t , and let $\mathcal{S} = \bigcup_t \mathcal{S}_t$. We can think of S_t as capturing the *state* of a deterministic algorithm at time t . For now, it is fine to think of S_t as interchangeable with K_t ; i.e., the knowledge set K_t captures all relevant details about all feedback the algorithm has observed thus far. (Later, when looking at L^p similarity metrics, we will want to keep track of separate knowledge sets at different scales, and thus will want a more nuanced notion of potential-based algorithm).

Definition 9. A deterministic algorithm \mathcal{A} (for $k = 2$ and $\alpha > 0$) is a potential-based algorithm if there exists a potential function Φ from \mathcal{S} to $\mathbb{R}_{\geq 0}$ and a “loss bound” function L from $\mathcal{S} \times B_d$ to $\mathbb{R}_{\geq 0}$ that satisfy:

- For all rounds t , $\Phi(S_{t+1}) \leq \Phi(S_t)$.
- Let q_t be the query point in round t . Then $L(S_t, q_t)$ is an upper bound on the loss incurred by **any guess**. In other words, $L(S_t, q_t)$ must satisfy

$$L(S_t, q_t) \geq \max_{\substack{x_1, x_2 \\ \text{consistent with } S_t}} |\delta(q_t, x_1) - \delta(q_t, x_2)|^\alpha.$$

- Again, let q_t be the query point in round t . If \mathcal{A} guesses the label incorrectly in round t , then $\Phi(S_t) - \Phi(S_{t+1}) \geq L(S_t, q_t)$.

For the case of inner-product similarity, we will set the loss bound function $L(S_t, q_t)$ equal to the width of the knowledge set K_t in the direction q_t . Importantly, this choice of L is an efficiently computable (in terms of q_t and the knowledge set K_t) upper-bound on the loss, which will prove important in the following section (in general, we will want both Φ and L to be efficiently computable in order to efficiently carry out the reduction in Section 3.2).

Note also that such a potential immediately gives a way to bound the total loss of an algorithm independently of T ; in particular, summing the inequality $\Phi(S_t) - \Phi(S_{t+1}) \geq L(S_t, q_t)$ over all t gives that the total loss is at most $\Phi(S_0)$. We call the value $\Phi(S_0)$ the *initial potential* of the algorithm \mathcal{A} .

Similar potential-based arguments are used in [LLS20] and [LS18] to give T -independent total loss bounds for the problem of contextual search. Unsurprisingly, these arguments can be extended (via Theorem 8) to apply to the problem of learning nearest neighbors as well.

Theorem 10. Fix an $\alpha > 0$. When $k = 2$, there exists a potential-based algorithm for learning nearest neighbor partitions under the similarity metric $\delta(x, y) = -\langle x, y \rangle$ that incurs total loss at most $O(\alpha^{-2} d \log d)$ (independent of the time horizon T).

The proof of Theorem 10 can be found in Appendix B.

3.2 From Two to Many Centers

We will now show how to use any potential-based algorithm for learning nearest neighbor partitions with two centers to construct an algorithm that can learn nearest neighbor partitions with any number of centers.

Our main result is the following:

Theorem 11. Let \mathcal{A} be a potential-based algorithm for learning nearest neighbor partitions with two centers that has an initial potential (and thus a total loss) of at most R . Then there exists a randomized algorithm \mathcal{A}' for learning nearest neighbor partitions for any $k \geq 2$ whose total expected loss is at most $O(k^2 R)$.

Similar to many existing methods for multiclass classification (“all-to-all” methods), we will accomplish this by running one instance of our two-center algorithm \mathcal{A} for each of the $\binom{k}{2}$ pairs of centers.

However, instead of using a simple majority voting scheme to choose our eventual label, we will use the *potentials* of these $\binom{k}{2}$ algorithms to construct a distribution over centers that we will sample from.

More specifically, our algorithm will work as follows. As mentioned, each round, based on the current query q_t and the potentials of the $\binom{k}{2}$ sub-algorithms, we will construct a distribution $v \in \Delta([k])$ over the k centers (we will describe how we do this shortly). We then sample a label i from this distribution v and guess it as the label of q_t . If we then learn that the correct label was in fact j , we update the sub-algorithm for the pair (i, j) with this information. We do not update any of the other sub-algorithms (in particular, if we guess the label correctly, we do not update any of the sub-algorithms).

To construct our distribution v , we will choose a distribution v that has the property that our expected loss in each round is at most the expected decrease in the total potential over all $\binom{k}{2}$ sub-algorithms. This will guarantee that the total expected loss of our algorithm is bounded above by the total starting potential of all our $\binom{k}{2}$ sub-algorithms. To be more precise, define the following variables:

1. Let \mathcal{A}_{ij} denote the two-center sub-algorithm for the labels i and j . Let $S_{ij}^{(t)}$ be the state of \mathcal{A}_{ij} at round t , and let $\Phi_{ij}^{(t)} = \Phi(S_{ij}^{(t)})$ be the potential of \mathcal{A}_{ij} at round t . Define $\Phi^{(t)} = \sum_{(i,j)} \Phi_{ij}^{(t)}$ to be the total of all the potentials belonging to sub-algorithms \mathcal{A}_{ij} .
2. As in Definition 9, let $L_{ij}^{(t)} = L(S_{ij}^{(t)}, q_t)$ denote an upper-bound on the loss incurred by the algorithm \mathcal{A}_{ij} in round t .
3. Let $D_{ij}^{(t)}$ denote the reduction in the potential of \mathcal{A}_{ij} when i is the correct label. In other words, $D_{ij}^{(t)} = \Phi_{ij}^{(t)} - \Phi_{ij}^{(t+1)}$ when $I_t^* = i$. Note that $D_{ij}^{(t)}$ is *not* equal to $D_{ji}^{(t)}$; it is possible for the potential of \mathcal{A}_{ij} to decrease a lot more upon learning that a point q_t has label i than learning it has label j (geometrically, this corresponds to different halves of the knowledge set being maintained by \mathcal{A}_{ij}).
4. Finally, define $M^{(t)} \triangleq D^{(t)} - \frac{1}{2} \cdot L^{(t)}$, where here we are treating $M^{(t)}$ and $L^{(t)}$ as n -by- n matrices. Observe that we can efficiently compute the values $L_{ij}^{(t)}$ and $D_{ij}^{(t)}$ and hence the value of $M_{ij}^{(t)}$ from q_t and the knowledge set of \mathcal{A}_{ij} at the beginning of round t .

From here on, we will fix a round t and suppress all associated superscripts. Assume that in this round the correct label for the query point is r . Now, if we sample a label from a distribution v , then note that the expected loss we sustain is $\sum_i L_{ri} v_i = e_r^T L v$. Similarly, the expected decrease in Φ is $\sum_i D_{ri} v_i = e_r^T D v$. If it was guaranteed to be the case that $e_r^T D v \geq e_r^T L v$, then this would in turn guarantee that our total expected loss is at most the total starting potential.

It follows that if we can find a distribution v that satisfies $M v \geq 0$, we are in luck. If such a distribution exists, we can find it by solving an LP. This is in fact how we find the distribution v , and this concludes the description of the algorithm. To prove correctness of the algorithm, it suffices to show that such a distribution always exists.

To do so, note that from the third point in Definition 9, we know that for each pair of labels (i, j) ,

$$D_{ij} + D_{ji} \geq L_{ij} \tag{P1}$$

(In fact, Definition 9 tells us that $\max(D_{ij}, D_{ji}) \geq L_{ij}$, since if \mathcal{A}_{ij} predicts j , we have that $D_{ij} \geq L_{ij}$, and likewise if \mathcal{A}_{ij} predicts i , we have that $D_{ji} \geq L_{ij}$). We can rewrite (P1) in the form $M + M^T \geq 0$. The following lemma shows that if $M + M^T \geq 0$, then there must exist a distribution v satisfying $M v \geq 0$, whose proof we defer to the appendix.

Lemma 3.2. *Given any matrix $M \in \mathbb{R}^{n \times n}$ such that $M + M^T \geq 0$, there exists a point $v \in \Delta_n$ such that $M v \geq 0$.*

With this, it is straightforward to finish off the proof of Theorem 11.

Proof of Theorem 11. By Lemma 3.2, we know that each round we can find a distribution $v \in \Delta([k])$ satisfying $M v \geq 0$. By the previous discussion, it follows that if we always sample from this

distribution, the total expected loss will be at most twice the starting total potential, $2\Phi^{(1)}$. But note that $\Phi^{(1)}$ is just the sum of the starting potentials $\Phi_{ij}^{(1)}$ of all the sub-algorithms, and is thus at most $\binom{k}{2}R$. It follows that the total loss of our new algorithm is at most $O(k^2R)$. \square

Corollary 3.3. *Fix an $\alpha > 0$. There exists a randomized algorithm for learning nearest neighbor partitions with the inner-product similarity metric that incurs total loss at most $O(\alpha^{-2}k^2d \log d)$.*

Remark 3.4. *Why do simple algorithms (such as a majority voting scheme) fail to work in our setting? In fact, it is possible to get a simple majority vote (breaking ties arbitrarily) to work if we are given additional feedback from the algorithm – specifically, the ranking of all k distances $\delta(q_t, x_i)$. With this information, it is possible to update all $\binom{k}{2}$ sub-instances each round, and charge any regret we sustain to an appropriate sub-instance. But if we only receive the true label of q_t , we no longer have the information to update every sub-instance, and instead have to do the more subtle amortization described above.*

4 Learning Nearest Neighbor Partitions Under L^p Similarity

In this section, we will discuss generalizations of our previous results for the inner-product similarity metric to general L^p spaces. We will primarily deal with the case when there are only two unknown points ($k = 2$) as the general reduction in Section 3.2 will allow us to reduce from the k -point case to the 2-point case.

The general approach for the algorithms in this section is to apply some sort of kernel mapping so that inequalities of the form $\|X - x_1\|_p \leq \|X - x_2\|_p$ become linear constraints. Once we linearize the problem, we can apply our earlier algorithms for the inner-product similarity along with tools from [LLS20].

Similar to the previous sections, we will assume that the hidden points x_1, x_2 are in the L^2 unit ball B_d . This is equivalent to assuming that the hidden points are in the L^p unit ball (which may be a more natural setting since we are working with L^p distances) up to a \sqrt{d} factor since we can simply rescale the L^2 -ball to contain the L^p -ball.

4.1 p -Norms for Even Integers p

When p is an even integer, there is a kernel mapping that exactly linearizes the problem. To see this, note that $|(x - a)|^p = (x - a)^p$ which is a polynomial in x so it suffices to consider the polynomial kernel $(1, x, \dots, x^p)$. In d dimensions, we can simply apply this kernel map coordinate-wise. After applying these kernel maps, we will be able to apply Corollary 3.3. Our main theorem for even integer p is stated below.

Theorem 12. *For even integer p , there is an algorithm for learning nearest neighbor partitions under the L^p similarity metric that incurs expected total loss at most $O(p^4 d^{(p+1)/p} k^2 \log d) = O(k^2 \cdot \text{poly}(p, d))$.*

The details of the proof are deferred to Appendix E.2.

Remark 4.1. *For the special case of $p = 2$, the reduction is even more immediate (the kernel map needs only add a single dimension), and obtains a slightly tighter bound of $O(k^2 d \log d)$. The reduction for this special case is summarized in Appendix C.*

4.2 General p -Norms

Now we discuss how to deal with general L^p norms. The main difficulty here is that there is no kernel map that exactly linearizes the problem so instead we will have multiple kernel maps. These kernel maps approximately linearize the problem at different scales, i.e. they have different output dimensions and as the output dimension grows, the problem can be more closely approximated by a linear one. When we are given a query point, we choose the scale of approximation that we use based on estimating the maximum possible loss that we can incur. By balancing the dimensionality and the approximation error to be at the same scale as the maximum possible loss, we may ensure that whenever we guess the label incorrectly, a certain potential function must decrease by an amount comparable to the loss that we incur. Our main theorem is stated below.

Theorem 13. Fix a $p > 2$. If all k unknown centers are Δ -separated in L^p distance, there exists an algorithm for learning nearest neighbor partitions under the L^p similarity metric that incurs total loss

$$\frac{k^2 \text{poly}(d, p)}{\Delta} \cdot \left(\frac{1}{p-2} \right)^2.$$

The full proof is more complicated than the algorithms in previous sections and requires opening the contextual search black-box and redoing parts of the analysis. Note that in the above theorem, we need the assumption that the unknown centers are Δ -separated, an assumption that was not necessary for even integer p . This is due to the fact that when the true centers are too close together, the dimensionality of the kernels that we need to achieve the necessary approximations are large. Nevertheless, we believe that the separated centers assumption is realistic for classification problems in practice. The details of the proof of Theorem 13 are deferred to Appendix E.3 of the Supplementary Material.

5 Learning General Convex Regions: Lower Bound

In this section, we consider the task of learning general convex regions, and present a construction which shows that any learning algorithm incurs $\Omega(T^{(d-4)/(d-2)})$ error over T rounds, even for only $k = 2$ regions. We give the full proof in Appendix F.

Theorem 4. Any algorithm for learning general convex regions incurs a total loss of at least $\Omega(T^{(d-4)/(d-2)})$, even when there are only two regions.

References

- [BR20] Aditya Bhaskara and Aravinda Kanchana Rwanpathirana. Robust algorithms for online k -means clustering. In *Algorithmic Learning Theory*, pages 148–173. PMLR, 2020.
- [CDK⁺06] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive aggressive algorithms. 2006.
- [CG13] Koby Crammer and Claudio Gentile. Multiclass classification with bandit feedback using adaptive regularization. *Machine learning*, 90(3):347–383, 2013.
- [CLPL16] Maxime C Cohen, Ilan Lobel, and Renato Paes Leme. Feature-based dynamic pricing. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 817–817. ACM, 2016.
- [CS03] Koby Crammer and Yoram Singer. Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, 3(Jan):951–991, 2003.
- [DH⁺73] Richard O Duda, Peter E Hart, et al. *Pattern classification and scene analysis*, volume 3. Wiley New York, 1973.
- [FS99] Yoav Freund and Robert E Schapire. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296, 1999.
- [GLK⁺17] Claudio Gentile, Shuai Li, Purushottam Kar, Alexandros Karatzoglou, Giovanni Zappella, and Evans Etrue. On context-dependent clustering of bandits. In *International Conference on Machine Learning*, pages 1253–1262. PMLR, 2017.
- [GLZ14] Claudio Gentile, Shuai Li, and Giovanni Zappella. Online clustering of bandits. In *International Conference on Machine Learning*, pages 757–765. PMLR, 2014.
- [HVD15] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.
- [Kle55] VL Klee. Separation properties of convex cones. *Proceedings of the American Mathematical Society*, 6(2):313–318, 1955.

- [KSST08] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Efficient bandit algorithms for online multiclass prediction. In *Proceedings of the 25th international conference on Machine learning*, pages 440–447, 2008.
- [Li11] Shengqiao Li. Concise formulas for the area and volume of a hyperspherical cap. *Asian Journal of Mathematics and Statistics*, 4(1):66–70, 2011.
- [LLS20] Allen Liu, Renato Paes Leme, and Jon Schneider. Optimal contextual pricing and extensions, 2020.
- [LLV17] Ilan Lobel, Renato Paes Leme, and Adrian Vladu. Multidimensional binary search for contextual decision-making. *Operations Research*, 2017.
- [LS18] Renato Paes Leme and Jon Schneider. Contextual search via intrinsic volumes. In *59th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2018, Paris, France, October 7-9, 2018*, pages 268–282, 2018.
- [LSS16] Edo Liberty, Ram Sriharsha, and Maxim Sviridenko. An algorithm for online k-means clustering. In *2016 Proceedings of the eighteenth workshop on algorithm engineering and experiments (ALENEX)*, pages 81–89. SIAM, 2016.
- [MG07] Jiri Matousek and Bernd Gärtner. *Understanding and using linear programming*. Springer Science & Business Media, 2007.
- [MR13] Mehryar Mohri and Afshin Rostamizadeh. Perceptron mistake bounds. *arXiv preprint arXiv:1305.0208*, 2013.
- [Nov63] Albert B Novikoff. On convergence proofs for perceptrons. Technical report, STANFORD RESEARCH INST MENLO PARK CA, 1963.
- [Ros58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [SBD06] Nathan Srebro and Shai Ben-David. Learning bounds for support vector machines with learned kernels. In *International Conference on Computational Learning Theory*, pages 169–183. Springer, 2006.
- [Vap13] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

Checklist

1. For all authors...
 - (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [\[Yes\]](#)
 - (b) Did you describe the limitations of your work? [\[Yes\]](#)
 - (c) Did you discuss any potential negative societal impacts of your work? [\[N/A\]](#)
 - (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [\[Yes\]](#)
2. If you are including theoretical results...
 - (a) Did you state the full set of assumptions of all theoretical results? [\[Yes\]](#)
 - (b) Did you include complete proofs of all theoretical results? [\[Yes\]](#)
3. If you ran experiments...
 - (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [\[N/A\]](#)
 - (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [\[N/A\]](#)
 - (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [\[N/A\]](#)

- (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
- 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
 - (a) If your work uses existing assets, did you cite the creators? [N/A]
 - (b) Did you mention the license of the assets? [N/A]
 - (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
 - (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
 - (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
- 5. If you used crowdsourcing or conducted research with human subjects...
 - (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
 - (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
 - (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

Funding Transparency Statement

Allen Liu is supported by a NSF Graduate Fellowship and a Hertz fellowship. None of the other authors received third-party funding or support during the preparation of this paper.

A Contextual search

In this appendix, we review the contextual search algorithm and analysis presented in [LLS20]. Our presentation will largely follow that of [LLS20], with two minor changes: 1. we will demonstrate that the algorithm works for all loss functions of the form $\ell(g_t, \langle x_t, p \rangle)$, where it incurs total loss at most $O(\alpha^{-2} d \log d)$ (Theorem 7), and 2. we will present the analysis in a slightly different way that makes it easier for us to construct potential-based algorithms for learning nearest neighbors.

The algorithm is outlined in Algorithm 1. Briefly, the algorithm works as follows. Whenever the algorithm gets a query direction x_t from the adversary, the algorithm looks at the width of the current knowledge set in the direction x_t . Based on the size of this width, the algorithm picks an “expansion parameter” z_i , and chooses a guess g_t so that the hyperplane $\langle v, x_t \rangle = g_t$ splits the volume of $K_t + z_i B_d$ in half.

Algorithm 1 CONTEXTUAL SEARCH ALGORITHM ([LLS20])

Initialize $K_1 = B_d$ and $z_i = 2^{-i}/(8d)$ for all i .
for t in $1, 2, \dots, T$ **do**
 Adversary picks x_t .
 Let i be the largest index such that $\text{width}(K_t; x_t) \leq 2^{-i}$.
 Submit guess g_t such that $\text{Vol}(\{v \in K_t + z_i B_d \mid \langle v, x_t \rangle \geq g_t\}) = \frac{1}{2} \text{Vol}(K_t + z_i B_d)$.
 Update K_{t+1} based on feedback.

Theorem 14 (Restatement of Theorem 7). *Let $\alpha > 0$. Algorithm 1 is an algorithm for contextual search with loss function $\ell(g_t, \langle x_t, p \rangle) = |g_t - \langle x_t, p \rangle|^\alpha$ that incurs a total loss of at most $O(\alpha^{-2} d \log d)$.*

Proof. To prove Theorem 7, we will examine the following potential function of the knowledge set at time t .

$$\Phi(K_t) = \sum_{i=1}^{\infty} 2^{-\alpha i} \log \frac{\text{Vol}(K_t + z_i B_d)}{\text{Vol}(z_i B_d)}.$$

Our goal will be to show that $\Phi(K_t)$ decreases by at least the loss we sustain in each round. This will bound the total loss Algorithm 1 sustains by at most $\Phi(K_0)$. To do this, we will employ the following lemma from [LLS20]:

Lemma A.1 (Lemma 2.1 in [LLS20]). *If i is the index chosen at round t in Algorithm 1, then $\text{Vol}(K_{t+1} + z_i B_d) \leq \frac{3}{4} \text{Vol}(K_t + z_i B_d)$.*

Note that Lemma A.1 implies that if i is the index chosen at round t , then $\Phi(K_t) - \Phi(K_{t+1}) \geq (\log \frac{4}{3}) 2^{-\alpha i}$. But also, if i is the chosen index at round t , then the width in the query direction is at most 2^{-i} , and thus the loss sustained in this round is at most $2^{-\alpha i}$. If we let L_t be the loss sustained in round t , we have thus shown that

$$\Phi(K_t) - \Phi(K_{t+1}) \geq \left(\log \frac{4}{3} \right) L_t.$$

Summing this over all t , we find the total loss is at most $O(\Phi(K_1))$. Since $K_1 = B_d$, we can evaluate $\Phi(K_1)$ as follows:

$$\begin{aligned}
\Phi(K_1) &= \sum_{i=1}^{\infty} 2^{-\alpha i} \log \frac{\text{Vol}(B_d + z_i B_d)}{\text{Vol}(z_i B_d)} \\
&= \sum_{i=1}^{\infty} 2^{-\alpha i} d \log \left(1 + \frac{1}{z_i} \right) \\
&\leq \sum_{i=1}^{\infty} 2^{-\alpha i} d \log (2^{i+4} d) \\
&= O \left(d \sum_{i=1}^{\infty} 2^{-\alpha i} i \right) + O \left(d \log d \sum_{i=1}^{\infty} 2^{-\alpha i} \right) \\
&= O(\alpha^{-2} d) + O(\alpha^{-1} d \log d) \\
&\leq O(\alpha^{-2} d \log d).
\end{aligned}$$

□

B Potential-based algorithm for inner-product similarity

In this appendix we prove Theorem 10, showing that the described algorithm for learning nearest neighbor partitions in Section 3.1 is a potential-based algorithm. Indeed, we will be able to use the same potential function as in the proof of Theorem 14, namely:

$$\Phi(K_t) = \sum_{i=1}^{\infty} 2^{-\alpha i} \log \frac{\text{Vol}(K_t + z_i B_d)}{\text{Vol}(z_i B_d)}.$$

Note that $\Phi(K_t)$ clearly satisfies the first two conditions in Definition 9, namely $\Phi(K_t) \geq 0$ for any knowledge set K_t , and it is always the case that $\Phi(K_{t+1}) \leq \Phi(K_t)$ (in particular, since $K_{t+1} \subseteq K_t$). It thus suffices to show the third condition of Definition 9 holds:

Lemma B.1. *Let \mathcal{A} be the ($k = 2$) algorithm for learning nearest neighbor partitions with similarity metric $\delta(x, y) = -\langle x, y \rangle$ and loss raised to the power α described in Corollary 3.1. Then, for all rounds t where \mathcal{A} guesses the label incorrectly, $\Phi(K_t) - \Phi(K_{t+1}) \geq \Omega(\max_{v \in K_t} |\langle q_t, v \rangle|^\alpha)$.*

Proof. Recall from the proof of Theorem 8, whenever \mathcal{A} guesses the label incorrectly, we update the state of the contextual search algorithm underlying \mathcal{A} . The contextual search algorithm underlying \mathcal{A} shares the same knowledge set K_t as \mathcal{A} , and by the analysis in the proof of Theorem 14, $\Phi(K_t)$ must then satisfy

$$\Phi(K_t) - \Phi(K_{t+1}) \geq \left(\log \frac{4}{3} \right) \text{width}(K_t; q_t)^\alpha.$$

Since $\text{width}(K_t; q_t) = \max_{v \in K_t} \langle v, q_t \rangle - \min_{v \in K_t} \langle v, q_t \rangle$, $\text{width}(K_t; q_t) \geq \max_{v \in K_t} |\langle q_t, v \rangle|$, and we have proved this lemma. □

C From Euclidean Distance to Inner-Product Similarity

We have presented two different variants of the nearest neighbor partition problem: one where we want to return the point x_i with largest inner-product similarity to each query q_t , and one where we want to return the point x_i closest to the query q_t in some L^p norm. Here we will show that in the case of the Euclidean norm, we can easily reduce the second problem to the first – and therefore, it suffices to solve the problem only for the case of inner-product similarity in Section 3. More specifically, we will show that if we can solve the nearest neighbor partition problem for the similarity metric $\delta(x, y) = -\langle x, y \rangle$ and $\alpha = 1/2$, we can solve the nearest neighbor partition problem for the L^2 similarity metric $\delta(x, y) = \|x - y\|_2$. (In many ways, this can be seen as a warm-up for the more general case of even integer p L^p norms in Appendix E.2).

Let x_1, x_2, \dots, x_k be points in B_d . Consider following two maps T and Q from $B_d \rightarrow B_{d+1}$. T maps the point $x \in B_d$ to $T(x) \triangleq \frac{1}{\sqrt{2}}(x, \|x\|_2^2)$ where $(x, \|x\|_2^2)$ is the $(d+1)$ -dimensional vector formed by appending $\|x\|_2$ to x . Q maps the point $q \in B_d$ to $Q(q) \triangleq \frac{1}{\sqrt{5}}(2q, -1)$. We now have the following two claims.

Lemma C.1. *Let $X = \{x_1, x_2, \dots, x_k\}$ be a set of points in B_d . Then for any $q \in B_d$, if $x^* = \arg \min_{x \in X} \|q - x\|_2$, it is also true that $x^* = \arg \max_{x \in X} \langle T(x), Q(q) \rangle$.*

Proof. Consider two points $x, x' \in B_d$. It suffices to show that if $\|q - x\| \leq \|q - x'\|$, then $\langle T(x), Q(q) \rangle \geq \langle T(x'), Q(q) \rangle$.

To see this, note that we can rewrite $\|q - x\|^2 \leq \|q - x'\|^2$ in the form $\langle q - x, q - x \rangle \leq \langle q - x', q - x' \rangle$, which we can in turn simplify to get

$$-2 \langle q, x \rangle + \|x\|^2 \leq -2 \langle q, x' \rangle + \|x'\|^2. \quad (1)$$

But the LHS of (1) is simply $-\sqrt{10} \langle T(x), Q(q) \rangle$ while the RHS of (1) is likewise $-\sqrt{10} \langle T(x'), Q(q) \rangle$. Equation (1) thus implies that $\langle T(x), Q(q) \rangle \geq \langle T(x'), Q(q) \rangle$, as desired. \square

Lemma C.2. *Let x, x' , and q be points in B_d . Let*

$$\ell_1 \triangleq \left| \|q - x\| - \|q - x'\| \right|$$

and

$$\ell_2 \triangleq \left| \langle T(x), Q(q) \rangle - \langle T(x'), Q(q) \rangle \right|.$$

Then $\ell_1 \leq 2\sqrt{\ell_2}$.

Proof. Note that $\ell_1 \leq \|q - x\| - \|q - x'\|$, so in particular

$$\ell_1^2 \leq \left| \|q - x\|^2 - \|q - x'\|^2 \right|. \quad (2)$$

Via the same logic in the proof of Lemma C.1, we can rewrite the RHS of (2) as $\sqrt{10}\ell_2$. It follows that $\ell_1^2 \leq \sqrt{10}\ell_2$ and thus that $\ell_1 \leq 2\sqrt{\ell_2}$. \square

With these two lemmas, we can prove the following reduction.

Theorem 15. *Let \mathcal{A} be an algorithm for learning nearest-neighbor partitions under the similarity metric $\delta(x, y) = -\langle x, y \rangle$ with $\alpha = 1/2$ that achieves a total loss of at most $R(k, d)$. Then there exists an algorithm \mathcal{A}' for learning nearest-neighbor partitions under the similarity metric $\delta(x, y) = \|x - y\|_2$ (and $\alpha = 1$) that achieves a total loss of at most $2R(k, d + 1)$.*

Proof. To construct algorithm \mathcal{A}' from algorithm \mathcal{A} , we simply map each incoming query $q_t \in B_d$ for algorithm \mathcal{A}' to the point $q'_t = Q(q_t) \in B_{d+1}$ and feed it to \mathcal{A} (returning the label that \mathcal{A} outputs, and providing \mathcal{A} with the true label that we receive).

To see why this works, note that if the hidden centers for \mathcal{A}' are the points $x_1, x_2, \dots, x_k \in B_d$, then by Lemma C.1, all feedback we provide \mathcal{A} is consistent with the set of hidden centers $T(x_1), T(x_2), \dots, T(x_k) \in B_{d+1}$. Moreover, by Lemma C.2, whenever algorithm \mathcal{A} incurs loss ℓ , our algorithm \mathcal{A}' incurs loss at most 2ℓ . It follows that \mathcal{A}' incurs loss at most $2R(k, d + 1)$. \square

D Mistake Bounds

In this section we provide mistake bounds for our algorithms for learning linear classifiers and learning nearest neighbor partitions. In both cases we will get (near) state-of-the-art guarantees for the mistake bound, despite our algorithms being designed for the absolute loss function.

We begin by discussing our algorithm for learning linear classifiers. As noted in the introduction, note that since this algorithm incurs total loss of at most $O(d \log d)$, then we make at most $O(d \log d / \gamma)$ mistakes in a setting with margin γ . In the following theorem, we see that we can improve this bound to $O(d \log 1/\gamma + d \log d)$ (matching the mistake bound of the best halving-based algorithms whenever $\gamma \leq 1/d$).

Theorem 16. Assume every query point q_t we are provided satisfies $|\langle q_t, x_1 - x_2 \rangle| \geq \gamma$ for some $\gamma > 0$. Then the algorithm of Theorem 1 makes at most $O(d \log 1/\gamma + d \log d)$ mistakes.

Proof. The main observation is that since we query our contextual search subroutine (which is trying to learn the hidden point $w = x_1 - x_2$) with the point q_t , if $|\langle q_t, x_1 - x_2 \rangle| \geq \gamma$, then either 1. we already know for certain the sign of $\langle q_t, x_1 - x_2 \rangle$, or 2. the width $\text{width}(K_t; q_t) \geq \gamma$.

In the first case, we cannot make a mistake. In the second case, since the width is at least γ , it suffices to only consider the $\lceil \log(1/\gamma) \rceil$ th term of the potential function in Theorem 14. Specifically, note that by Lemma A.1, if we let $i = \lceil \log(1/\gamma) \rceil$ then we have that

$$\text{Vol}(K_{t+1} + z_i B_d) \leq \frac{3}{4} \text{Vol}(K_t + z_i B_d).$$

In particular, since $z_i = 2^{-i}/8d$, our total number of errors is at most

$$\log_{4/3} \frac{\text{Vol}(B_d + z_i B_d)}{\text{Vol}(z_i B_d)} = d \log_{4/3} \left(1 + \frac{1}{z_i} \right) = O(d \log(1/\gamma) + d \log d).$$

□

The same logic extends to learning nearest-neighbor partitions via the reduction in Theorem 11.

Theorem 17. Assume every query point q_t satisfies $\delta(q_t, R_i) > \gamma$ for all i such that $q_t \notin R_i$. Then the algorithm of Theorem 2 makes at most $O(k^2 d(\log 1/\gamma + \log d))$ mistakes.

Proof. We apply the reduction of Theorem 11 to Theorem 16. In particular, note that the analysis of Theorem 16 implies that our original algorithm for learning linear classifiers can be thought of (under these margin conditions) as a potential-based algorithm with loss function $L(S_t, q_t) = 1$ and with potential function

$$\Phi(S_t) = \log_{4/3} \frac{\text{Vol}((1 + z_i)B_d)}{\text{Vol}(S_t + z_i B_d)}.$$

The analysis of Theorem 16 combined with the guarantees of Theorem 11 imply a mistake bound of $O(k^2 d(\log 1/\gamma + \log d))$. □

Finally, we prove a general reduction for the notion of robust mistake bound defined in the introduction. Formally, the robust mistake bound with margin γ is the loss induced by the loss function $\ell'(q, R_i) = \mathbf{1}(\delta(q, R_i) - \delta(q, R^*) \geq \gamma)$ (where R^* is the region containing q). In the below lemma, we relate this to the loss induced by our standard loss function $\ell(q, R_i) = (\delta(q, R_i) - \delta(q, R^*))$.

Lemma D.1. If an algorithm has total loss at most R under the loss function $\ell(q, R_i) = (\delta(q, R_i) - \delta(q, R^*))$, it has a robust mistake bound of $O(R/\gamma)$ under margin γ .

Proof. This immediately follows from the fact that:

$$\ell(q, R_i) = (\delta(q, R_i) - \delta(q, R^*)) \geq \gamma \cdot \mathbf{1}(\delta(q, R_i) - \delta(q, R^*) \geq \gamma) = \gamma \ell'(q, R_i).$$

□

E Omitted Proofs

E.1 Omitted Proofs from Section 3.2

Proof of Theorem 8. The general idea behind this reduction is simple. We will run our algorithm \mathcal{A} for contextual search to find the hidden point $w = x_1 - x_2$. Whenever we are given a query point q_t and need to guess the sign of $\langle q_t, w \rangle$, we will ask our contextual search algorithm \mathcal{A} for a guess g_t for the value of $\langle q_t, w \rangle$. If $g_t > 0$, we will guess that $\langle q_t, w \rangle > 0$; otherwise, we will guess that $g_t < 0$.

There is one important caveat here: how do we update the algorithm \mathcal{A} ? Recall that any contextual search algorithm expects binary feedback each round as to whether its guess g_t was too high or too low. While in some cases, we can provide \mathcal{A} with accurate feedback, in many cases we cannot: for example, if \mathcal{A} submits a guess $g_t = 2.5$ for $\langle q_t, w \rangle$, and all we learn is that $\langle q_t, w \rangle \geq 0$, we cannot say with confidence whether \mathcal{A} 's guess was too large or not.

The solution to this is to *only update the state of \mathcal{A} on rounds where we guess the sign incorrectly*. Note that for such rounds, we definitively know whether g_t was too high or too low; for example, if \mathcal{A} guessed $g_t = 2.5$ and hence we guess $\langle q_t, w \rangle > 0$, but it turns out that $\langle q_t, w \rangle < 0$, we know for certain that the guess g_t was too high. On all other rounds we do not update the state of \mathcal{A} , effectively rolling back the state of \mathcal{A} to before we asked the question about q_t . This means that the effective number of rounds \mathcal{A} experiences (gets feedback on) may be less than T ; nonetheless, since $R(d, T)$ is non-decreasing in T , the total loss of \mathcal{A} on these rounds is still at most $R(d, T)$.

Finally, we will relate the loss of our algorithm \mathcal{A}' for learning nearest neighbors to the loss of the contextual search algorithm \mathcal{A} . To start, note that we only sustain loss in rounds when we guess the sign of $\langle q_t, w \rangle$ incorrectly. Luckily, these rounds happen to be exactly the rounds where we update the state of \mathcal{A} (and thus the rounds whose loss counts towards the $R(d, T)$ bound). In a round where we guess the sign incorrectly, \mathcal{A}' sustains a loss of $|\langle q_t, w \rangle|^\alpha$, and \mathcal{A} sustains a loss of $|\langle q_t, w \rangle - g_t|^\alpha$. Since $\text{sign}(g_t) \neq \text{sign}(\langle q_t, w \rangle)$, this means that $|\langle q_t, w \rangle - g_t| \geq |\langle q_t, w \rangle|$, and therefore \mathcal{A} sustains more loss than \mathcal{A}' . It follows that the total loss sustained by \mathcal{A}' is at most the total loss sustained by \mathcal{A} on this set of rounds, which in turn is at most $R(d, T)$. \square

Proof of Lemma 3.2. To show the existence of such a distribution v , we will show that the following linear program has a solution.

$$\begin{aligned} \min & 0 \\ & Mv \geq 0 \\ & \sum_i v_i = 1 \\ & v \geq 0 \end{aligned}$$

If the above program has no solution then by the Strong Duality Theorem (see [MG07] section 6.1), we know that the dual program below is unbounded. (Since the dual has a feasible solution $y = 0, z = 0$, we know that it must be unbounded). In particular, for any value $z > 0$, there exists a corresponding solution to y to the dual program.

$$\begin{aligned} \max & z \\ & M^T y + z\mathbf{1} \leq 0 \\ & y \geq 0 \end{aligned}$$

Let y be any solution to the above dual with $z = 1$. Since $M + M^T$ has all non-negative entries, we know that $(M + M^T)y \geq 0$. Combining this with the fact that $M^T y \leq -\mathbf{1}$, we get that $My \geq 0$. This contradicts the fact that there is no solution to the primal since $\frac{y}{\sum_i y_i}$ is a feasible point. \square

E.2 Omitted Proofs from Section 4.1

Here we prove Theorem 12. Fix an even integer p . We define the following kernel maps.

Definition 18. Let $\text{Ker} : \mathbb{R} \rightarrow \mathbb{R}^{p+1}$ be the map defined by

$$\text{Ker}(x) \triangleq (1, x, x^2, \dots, x^p).$$

Definition 19. For a point $y = (y_1, \dots, y_d) \in B_d$, define the map $G : B_d \rightarrow B_{(p+1)d}$ as

$$G(y) \triangleq \frac{1}{\sqrt{pd}} \left(\text{Ker}\left(\frac{y_1}{p}\right), \dots, \text{Ker}\left(\frac{y_d}{p}\right) \right)$$

where above the outputs of $\text{Ker}(\cdot)$ are simply concatenated.

Definition 20. Let $F : \mathbb{R} \rightarrow \mathbb{R}^{p+1}$ be defined by

$$F(a) \triangleq \left(a^p, -\binom{p}{1} a^{p-1}, \dots, -\binom{p}{p-1} a, \binom{p}{p} \right)$$

Note that if $|a| \leq 1/p$, all components of $F(a)$ are at most 1 in absolute value.

Definition 21. For a point $z = (z_1, \dots, z_d) \in \mathbb{R}^d$, define the map $H : B_d \rightarrow B_{(p+1)d}$ as

$$H(z) \triangleq \frac{1}{\sqrt{pd}} \left(F\left(\frac{z_1}{p}\right), \dots, F\left(\frac{z_d}{p}\right) \right).$$

The key property that these maps satisfy is stated below.

Lemma E.1. For points $y, z \in \mathbb{R}^d$,

$$\langle G(y), H(z) \rangle = \frac{1}{pd} \left\| \frac{y-z}{p} \right\|_p^p.$$

Proof. The proof follows by substituting in the definitions for G and H and using the binomial theorem. \square

In particular, we can rewrite the statement of Lemma E.1 in the form

$$\|y-z\|_p = p^{(p+1)/p} d^{1/p} \langle G(y), H(z) \rangle^{1/p}. \quad (3)$$

This suggests a reduction to the inner-product similarity metric similar to the reduction for Euclidean norm in Theorem 15. In particular, note that for $p \geq 1$, we have that:

$$|x|^p - |y|^p \geq |x-y|^p.$$

In particular, this implies that

$$\|q-x\| - \|q-x^*\| \leq (\|q-x\|^p - \|q-x^*\|^p)^{1/p} = p^{(p+1)/p} d^{1/p} |\langle G(q), H(x^*) \rangle - \langle G(q), H(x) \rangle|^{1/p}.$$

Thus, if we use our map H to map each query point q_t to the point $H(q_t) \in B_{p(d+1)}$ and feed it into an algorithm with similarity metric $\delta(x, y) = -\langle x, y \rangle$ and $\alpha = 1/p$, Lemma E.1 implies that this algorithm will successfully learn the partition induced by the points $G(x_i)$. We can now complete the proof of Theorem 12.

Proof of Theorem 12. From Corollary 3.3, there exists an algorithm for learning nearest neighbor partitions with similarity function $\delta(x, y) = -\langle x, y \rangle$ and $\alpha = 1/p$ with expected total loss $O(p^2 k^2 d \log d)$. Applying this algorithm as described above (noting that the ambient dimension is now $p(d+1)$ and the loss is scaled by a factor of $O(pd^{1/p})$), we obtain the bound in the statement. \square

E.3 Omitted Proofs from Section 4.2

Here, we prove Theorem 13.

E.3.1 Kernelization

Fix the L^p norm that we are working with. Let $p' = \lfloor p \rfloor + 1$. For each $i = 1, 2, \dots$, let

$$\delta_i = \frac{1}{100d^2 p' 2^i}, D_i = \frac{1}{2\delta_i}.$$

We will now define two maps that will be crucial for our algorithm.

For each i , we will define two maps G_i, H_i that map points in \mathbb{R}^d to points in $\mathbb{R}^{p'd(2D_i+1)}$ such that for two points $y, z \in \mathbb{R}^d$, the images $G_i(y), H_i(z)$ satisfy the property that $\langle G_i(y), H_i(z) \rangle$ is a good approximation of $\|y-z\|_p^p$. We need to consider maps for different values of i because as i increases, the approximation gets better but the dimension of the image space also increases. We can think of these maps for different values of i as approximations at different scales.

Through the next several definitions, we build the first map G_i .

Definition 22. For $x \neq 0$, let $\text{sign}(x) \triangleq x/|x|$. Let $\text{sign}(0) \triangleq 0$.

Definition 23. Let $D : \mathbb{R} \rightarrow \mathbb{R}^{p'}$ be the map defined by

$$D(x) = (|x|^p, \mathbf{sign}(x) |x|^{p-1}, |x|^{p-2}, \mathbf{sign}(x) |x|^{p-3}, \dots, \mathbf{sign}(x)^{\lfloor p \rfloor} |x|^{p-\lfloor p \rfloor}).$$

Definition 24. Let $\mathbf{Ker}_i : \mathbb{R} \rightarrow \mathbb{R}^{p'(2D_i+1)}$ be the map defined by

$$\mathbf{Ker}_i(x) = (D(x + 0.5), D(x + 0.5 - \delta_i), \dots, D(x - 0.5))$$

where the tuples given by the output of $D(\cdot)$ are simply concatenated.

Definition 25. For a point $y = (y_1, \dots, y_d) \in B_d$, define the map $G_i : \mathbb{R}^d \rightarrow \mathbb{R}^{p'd(2D_i+1)}$ as

$$G_i(y) = (\mathbf{Ker}_i(y_1/2), \dots, \mathbf{Ker}_i(y_d/2)).$$

Now we build the second map H_i through the next set of definitions.

Definition 26. Let $F_i : [-1/2, 1/2] \rightarrow \mathbb{R}^{p'(2D_i+1)}$ be the map defined as follows. Assume that we want to compute $F_i(x)$. Then perform the following steps.

- Let c be the unique integer such that $c\delta_i \leq x < (c+1)\delta_i$
- For any element of $\mathbb{R}^{p'(2D_i+1)}$, group the coordinates into consecutive groups of p' and label the groups with $-D_i, -D_i + 1, \dots, D_i$
- Let $F_i(x)$ be the element of $\mathbb{R}^{p'(2D_i+1)}$ where
 - The group labeled c is set to

$$\left(1, p(x - c\delta_i), \frac{p(p-1)}{2}(x - c\delta_i)^2, \frac{p(p-1)(p-2)}{6}(x - c\delta_i)^3, \dots \right)$$

- All other groups are set to $(0, 0, \dots, 0)$

Definition 27. For a point $z = (z_1, \dots, z_d) \in B_d$, define the map $H_i : \mathbb{R}^d \rightarrow \mathbb{R}^{p'd(2D_i+1)}$ as

$$H_i(z) = (F_i(z_1/2), \dots, F_i(z_d/2)).$$

The intuition for the interplay between the maps G_i and H_i is that the “kernel” map G_i discretizes the function $|x|^p$ as well as its derivatives and then H_i takes the first p' terms of the Taylor series expansion at the closest point in the discretization.

Formally, the key property that the maps G_i, H_i satisfy is the following:

Lemma E.2. For points $y, z \in B_d$,

$$\left| \langle G_i(y), H_i(z) \rangle - \left\| \frac{y-z}{2} \right\|_p^p \right| \leq d(p\delta_i)^p$$

The proof of Lemma E.2 relies on the following inequality.

Claim 28. Let $p > 2$. Then for any $x, x' \in [-1, 1]$, we have the inequality

$$\left| |x|^p - \sum_{i=0}^{\lfloor p \rfloor} \frac{p(p-1)\dots(p-i+1)}{i!} (x-x')^i \mathbf{sign}(x')^i |x'|^{p-i} \right| \leq (p|x-x'|)^p.$$

Proof. Note that the function $|x|^p$ is $\lfloor p \rfloor$ -times continuously differentiable and its derivatives are $p \cdot \mathbf{sign}(x) |x|^{p-1}, p(p-1) |x|^{p-2}, \dots$ and so on. Thus, we may write

$$\begin{aligned} |x|^p &= |x'|^p + \int_{x'}^x p \cdot \mathbf{sign}(y) |y|^{p-1} dy \\ &= |x'|^p + p(x-x')\mathbf{sign}(x') |x'|^{p-1} + \int_{x'}^x \int_{x'}^{y_1} p(p-1) |y_2|^{p-2} dy_2 dy_1 \\ &\vdots \\ &= \sum_{i=0}^{\lfloor p \rfloor} \frac{p(p-1)\dots(p-i+1)}{i!} (x-x')^i \mathbf{sign}(x')^i |x'|^{p-i} \\ &\quad + \int_{x'}^x \dots \int_{x'}^{y_{\lfloor p \rfloor-1}} p \dots (p-\lfloor p \rfloor + 1) (\mathbf{sign}(y_{\lfloor p \rfloor})^{\lfloor p \rfloor} |y_{\lfloor p \rfloor}|^{p-\lfloor p \rfloor} - \mathbf{sign}(x')^{\lfloor p \rfloor} |x'|^{p-\lfloor p \rfloor}) dy_{\lfloor p \rfloor} \dots dy_1. \end{aligned}$$

It now suffices to bound the last term which is the “error” term.. However since $p - \lfloor p \rfloor < 1$,

$$\left| \text{sign}(y)^{\lfloor p \rfloor} |y|^{p-\lfloor p \rfloor} - \text{sign}(x')^{\lfloor p \rfloor} |x'|^{p-\lfloor p \rfloor} \right| \leq |y - x'|^{p-\lfloor p \rfloor},$$

so the error term is at most

$$p \cdots (p - \lfloor p \rfloor + 1) |x - x'|^p \leq (p|x - x'|)^p,$$

and now we immediately get the desired inequality. \square

Now we can prove Lemma E.2.

Proof of Lemma E.2. Let $y = (y_1, \dots, y_d)$ and $z = (z_1, \dots, z_d)$. For each $j \in [d]$ let c_j be the integer such that $c_j \delta_j \leq 0.5z_j < (c_j + 1)\delta_j$. Now we have

$$\langle G_i(y), H_i(z) \rangle = \sum_{j \in [d]} \left(\sum_{i=0}^{\lfloor p \rfloor} \frac{p(p-1) \cdots (p-i+1)}{i!} (0.5z_j - c_j \delta_i)^i \text{sign}(0.5y_j - c_j \delta_i)^i |0.5y_j - c_j \delta_i|^{p-i} \right).$$

Now we can apply Claim 28 with $x = 0.5(y_j - z_j)$ and $x' = 0.5y_j - c_j \delta_i$ to bound each term. Note that $|x - x'| < \delta_i$. Since the sum contains d terms, we immediately get the desired conclusion. \square

Our full algorithm for learning nearest neighbor partitions in L^p norm is described below.

E.3.2 Algorithm

Algorithm 2 MULTISCALE NEAREST NEIGHBOR LEARNING FOR L^p NORMS

There are two unknown points $x_1, x_2 \in B_d$

For each $i = 1, 2, \dots$ initialize the sets $S_i = [-1, 1]^{2p'd(2D_i+1)}$. Note that

$$S_i \supset \{(G_i(x), G_i(y)) | x, y \in [-1/2, 1/2]^d\}$$

for t in $1, 2, \dots, T$ **do**

Adversary picks $q_t \in B_d$

for $i = 1, 2, \dots$ **do**

Let $v_{i,t} = (-H_i(q_t), H_i(q_t))$

Let $w_{i,t}$ be the width of the set S_i in direction $v_{i,t}$ i.e.

$$w_{i,t} = \max_{u \in S_i} \langle v_{i,t}, u \rangle - \min_{u \in S_i} \langle v_{i,t}, u \rangle$$

Let i_t be the smallest integer i such that $w_{i,t} \geq 10^3 D_i p d^2 (p \delta_i)^p$

if $\text{Vol}(\{z \in S_{i_t}, v_{i_t,t} \cdot z > 0\}) \geq \text{Vol}(\{z \in S_{i_t}, v_{i_t,t} \cdot z < 0\})$ **then**

Guess label 1

if $\text{Vol}(\{z \in S_{i_t}, v_{i_t,t} \cdot z > 0\}) < \text{Vol}(\{z \in S_{i_t}, v_{i_t,t} \cdot z < 0\})$ **then**

Guess label 2

if true label is 1 **then** update

$$S_{i_t} \leftarrow \{z \in S_{i_t}, v_{i_t,t} \cdot z > -3d(p\delta_{i_t})^p\}$$

if true label is 2 **then** update

$$S_{i_t} \leftarrow \{z \in S_{i_t}, v_{i_t,t} \cdot z < 3d(p\delta_{i_t})^p\}$$

Remark E.3. In the algorithm, it is stated that we keep track of sets S_i for all integers i . Technically, this is not possible as there are infinitely many sets to keep track of but it will be clear from the analysis that it suffices to track the set S_i only for $i \leq \text{poly}(p, d, T)$ and if i_t is too large, we can simply guess arbitrarily and our loss will be upper bounded by $1/T$.

As in the previous section, for each timestep t and integer i , we let $S_i^{(t)}$ denote the set S_i at the beginning of timestep t in the execution of the algorithm.

Claim 29. For all i and all timesteps t , the set $S_i^{(t)}$ contains the L^2 ball of radius $0.1(p\delta_i)^p$ centered around $(G_i(x_1), G_i(x_2))$ where x_1, x_2 are the two unknown centers.

Proof. We will prove the claim by induction on t . The base case is obvious. Now we do the induction step. Consider a timestep t . Assume that the adversary gives us the point q_t . WLOG the true label of q_t is 1 i.e.

$$\|q_t - x_1\|_p^p \leq \|q_t - x_2\|_p^p .$$

Note that $\|v_{i,t}\|_2 \leq 10d$ for all i, t (this uses the fact that in Definition 26, $|x - c\delta_i| \leq \delta_i \leq 1/p$). Thus, any point z in the ball of radius $0.1(p\delta_i)^p$ centered around $(G_i(A), G_i(B))$ satisfies

$$|v_{i,t} \cdot z - v_{i,t} \cdot (G_i(x_1), G_i(x_2))| \leq d(p\delta_i)^p .$$

However

$$v_{i,t} \cdot (G_i(x_1), G_i(x_2)) = -\langle H_i(q_t), G_i(x_1) \rangle + \langle H_i(q_t), G_i(x_2) \rangle$$

and we can now use Lemma E.2 to deduce

$$\left| \left(-\left\| \frac{q_t - x_1}{2} \right\|_p^p + \left\| \frac{q_t - x_2}{2} \right\|_p^p \right) - v_{i,t} \cdot (G_i(x_1), G_i(x_2)) \right| \leq 2d(p\delta_i)^p .$$

Thus, by the triangle inequality, we must actually have for all z in the ball of radius $0.1(p\delta_i)^p$ centered around $(G_i(x_1), G_i(x_2))$,

$$v_{i,t} \cdot z \geq -3d(p\delta_i)^p$$

which implies that all of these points are all contained in S_i after the update step, completing the induction. \square

We will also need the following geometric fact.

Lemma E.4 (From [LLS20]). Let $S \subset \mathbb{R}^d$ be a convex polytope and v be a unit vector. Assume that the width of S in direction v is at least $8d\epsilon$. Then any strip of width ϵ normal to direction v contains at most $1/4$ of the volume of S .

Proof. Let C be a cross section of S normal to direction v with maximal area. Let u_1, u_2 be two points in S that minimize and maximize the inner product with v respectively. Then either u_1 or u_2 is distance at least $4d\epsilon$ from the hyperplane containing C . WLOG u_1 is at least $4d\epsilon$ away from this hyperplane. Then the cone containing u_1 and C must be contained in S (since S is convex) so

$$\text{Vol}(S) \geq 4\epsilon d \cdot \text{Vol}(C) \cdot \frac{1}{d} = 4\epsilon \text{Vol}C .$$

On the other hand, by the maximality of C , the volume contained in any ϵ -width strip is at most $\epsilon \text{Vol}(C)$ so we are done. \square

Claim 30. Consider a timestep t . If the learner incurs nonzero loss then

$$\text{Vol}\left(S_{i_t}^{(t+1)}\right) \leq \frac{3}{4} \text{Vol}\left(S_{i_t}^{(t)}\right) .$$

Proof. Without loss of generality, the true label is 1 and our guess was 2. Then we must have

$$\text{Vol}\left(\{z \in S_{i_t}^{(t)}, v_{i_t,t} \cdot z > 0\}\right) \leq \frac{1}{2} \text{Vol}\left(S_{i_t}^{(t)}\right) .$$

Also since $w_{i,t} \geq 10^3 D_{i_t} p d^2 (p\delta_{i_t})^p$ and the dimension of the space in which $S_{i_t}^{(t)}$ lives is $2p'd(2D_{i_t} + 1)$, Lemma E.4 gives us that

$$\text{Vol}\left(\{z \in S_{i_t}^{(t)}, -3d(p\delta_{i_t})^p < v_{i_t,t} \cdot z < 0\}\right) \leq \frac{1}{4} \text{Vol}\left(S_{i_t}^{(t)}\right) .$$

Thus, we deduce that

$$\text{Vol}(S_{i_t}^{(t+1)}) = \text{Vol}\left(\{z \in S_{i_t}^{(t)}, v_{i_t,t} \cdot z > -3d(p\delta_{i_t})^p\}\right) \leq \frac{3}{4} \text{Vol}\left(S_{i_t}^{(t)}\right) ,$$

as desired. \square

Theorem 31. *The total loss incurred by MULTISCALE NEAREST NEIGHBOR LEARNING is at most*

$$\frac{\text{poly}(d, p)}{\|x_1 - x_2\|_p} \cdot \left(\frac{1}{p-2}\right)^2.$$

Proof. Combining Claim 29 and Claim 30 implies that for any index i , the number of times that $i_t = i$ and we incur nonzero loss is at most

$$O\left(\log\left(\frac{2^{2p'd(2D_i+1)}}{(0.1(p\delta_i)^p)^{2p'd(2D_i+1)} \text{Vol}(B_{2p'(2D_i+1)}(0, 1))}\right)\right) = O(p^3 d^3 2^i (i + \log dp)).$$

Note that if $i_t = i$, then by the definition of our algorithm

$$w_{i-1,t} = \max_{u \in S_{i-1}^t} \langle v_{i-1,t}, u \rangle - \min_{u \in S_{i-1}^t} \langle v_{i-1,t}, u \rangle \leq 10^3 D_{i-1} p d^2 (p\delta_{i-1})^p.$$

Note that the origin is clearly always contained in S_{i-1} so using Claim 29, we get that

$$|\langle v_{i-1,t}, (G_{i-1}(x_1), G_{i-1}(x_2)) \rangle| \leq 10^3 D_{i-1} p d^2 (p\delta_{i-1})^p = O(d^2 p^2 (p\delta_{i-1})^{p-1}).$$

Lemma E.2 implies that

$$\left| \left(-\left\| \frac{q_t - x_1}{2} \right\|_p^p + \left\| \frac{q_t - x_2}{2} \right\|_p^p \right) - v_{i-1,t} \cdot (G_{i-1}(x_1), G_{i-1}(x_2)) \right| \leq 2d(p\delta_{i-1})^p,$$

so we deduce that

$$\left| \|q_t - x_1\|_p^p - \|q_t - x_2\|_p^p \right| \leq O(d^2 p^2 (4p\delta_{i-1})^{p-1}) \leq O(d^2 p^2 (10 \cdot 2^i d^2)^{-(p-1)}).$$

Note that our loss at each round may be bounded as

$$\left| \|q_t - x_1\|_p - \|q_t - x_2\|_p \right| \leq \frac{\left| \|q_t - x_1\|_p^p - \|q_t - x_2\|_p^p \right|}{\max(\|q_t - x_1\|_p, \|q_t - x_2\|_p)^{p-1}} \leq \frac{\left| \|q_t - x_1\|_p^p - \|q_t - x_2\|_p^p \right|}{\|0.5(x_1 - x_2)\|_p^{p-1}}.$$

Alternatively, we may also use the trivial bound

$$\left| \|q_t - x_1\|_p - \|q_t - x_2\|_p \right| \leq \|x_1 - x_2\|_p.$$

Let i_0 be the largest positive integer such that

$$\frac{i_0}{2^{i_0}} \geq 0.1 d^{-1} \|x_1 - x_2\|_p.$$

Note that $i_0 \geq 2$.

We can now bound the total loss of our algorithm, say L , as follows:

$$\begin{aligned} L &\leq \sum_{i=1}^{i_0} \|x_1 - x_2\|_p O(p^3 d^3 2^i (i + \log dp)) \\ &\quad + \sum_{i=i_0+1}^{\infty} \frac{\left| \|q_t - x_1\|_p^p - \|q_t - x_2\|_p^p \right|}{\|0.5(x_1 - x_2)\|_p^{p-1}} \cdot O(p^3 d^3 2^i (i + \log dp)) \\ &\leq \sum_{i=1}^{i_0} \|x_1 - x_2\|_p O(p^3 d^3 2^i (i + \log dp)) \\ &\quad + \frac{1}{\|0.5(x_1 - x_2)\|_p^{p-1}} \sum_{i=i_0+1}^{\infty} O(d^2 p^2 (10 \cdot 2^i d^2)^{-(p-1)}) \cdot O(p^3 d^3 2^i (i + \log dp)) \\ &\leq \text{poly}(d, p) \frac{1}{\|x_1 - x_2\|_p} \sum_{i=0}^{\infty} \frac{1+i}{2^{(p-2)i}} \\ &\leq \text{poly}(d, p) \cdot \frac{1}{\|x_1 - x_2\|_p} \cdot \frac{1}{(1 - 2^{-(p-2)})^2} \\ &\leq \frac{\text{poly}(d, p)}{\|x_1 - x_2\|_p} \cdot \left(\frac{1}{p-2}\right)^2. \end{aligned}$$

□

Proof of Theorem 13. In light of Theorem 11, it suffices to argue that MULTISCALE NEAREST NEIGHBOR LEARNING is a potential-based algorithm. Indeed, the corresponding potential is defined as follows. Let i_0 be the largest positive integer such that

$$\frac{i_0}{2^{i_0}} \geq 0.1d^{-1}\Delta.$$

Define

$$P_t = \sum_{i=1}^{i_0} \Delta \log \frac{\text{Vol}(S_i^{(t)})}{\text{Vol}(B_{2p^{i-1}}(0, 0.1 \cdot (p\delta_i)^p))} \\ + \sum_{i=i_0+1}^{\infty} \frac{d^2 p^2}{(10 \cdot 2^i d^2)^{p-1} (0.5\Delta)^{p-1}} \log \frac{\text{Vol}(S_i^{(t)})}{\text{Vol}(B_{2p^{i-1}}(0, 0.1 \cdot (p\delta_i)^p))}.$$

The proof of Theorem 31 immediately implies that P_t is a valid potential and we are done. \square

F Learning General Convex Regions: Missing Proofs

In this appendix, we present the full proof of Theorem 4.

Our construction is based on only picking points on the surface of the unit ball (i.e. the unit hypersphere). There are two key factors to ensuring that the algorithm accrues enough total error; we need to ensure that (i) each time we choose a point, it could lie in either of the two regions and (ii) the point is sufficiently far from the region it is not in. To guarantee the former, we choose our points by considering a separating hyperplane between the two regions so far.

To guarantee the latter, we will choose our points to be ϵ -far from each other, for some ϵ based on the total number of points we need to choose T . This, combined with the fact that we chose points on the surface of the unit ball, implies that the minimum penalty for a mistake is $\Omega(\epsilon^2)$. Note that for both of our guarantees to simultaneously work out, we actually need our separating hyperplane to go through the origin (so the resulting intersection has enough surface area for this part of the argument).

The requirement that points be ϵ far from each other limits the total number of points we can choose. Roughly speaking, each point removes on the order of a ϵ^{d-2} -fraction of the (hyper-)surface area of the intersection of the separating hyperplane with the unit hypersphere. Maximizing ϵ while ensuring we can pick T points yields the desired bound in the theorem statement.

Our construction will utilize two technical results concerning the geometry of high-dimensional objects. One is a result of Klee regarding the existence of separating hyperplanes for convex cones [Kle55]. It uses the following notation.

Definition 32 ([Kle55]). *A 0-cone is a closed convex cone having the origin (denoted 0) as its vertex. For a 0-cone A , A' denotes the linear subspace $A \cap -A$.*

The technical lemma gives conditions for a strict linear separator between two such convex cones.

Theorem 33 ([Kle55] Theorem 2.7). *Suppose E is a separable normed linear space, A and B are 0-cones in E , A is locally compact, and $A \cap B = \{0\}$. Then E admits a continuous linear functional \mathcal{H} such that $\mathcal{H} < 0$ on $A \setminus A'$, $\mathcal{H} = 0$ on $A' \cup B'$, and $\mathcal{H} > 0$ on $B \setminus B'$.*

The next technical lemma upper bounds the surface area of a hypersphere cap relative to the entire hypersphere, implying that many such caps are needed to cover the hypersphere.

Definition 34. *We denote the d -dimensional unit hypersphere as $\mathcal{S}_d \triangleq \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}$ (note that this is the surface of the unit ball). We denote its (hyper-)surface area as \mathcal{A}_d .*

Next, we denote the cap centered at $v \in \mathcal{S}_d$ of angle $\phi \in [0, \pi/2]$ as $\mathcal{S}_d(v, \phi) \triangleq \mathcal{S}_d \cap \{x \in \mathbb{R}^d \mid \langle x, v \rangle \leq \cos \phi\}$. We denote its (hyper-)surface area as $\mathcal{A}_d(\phi)$.

Lemma F.1. *For all integer $d \geq 1$ and $\phi \in [0, \pi/2]$,*

$$\frac{\mathcal{A}_d(\phi)}{\mathcal{A}_d} \leq \phi^{d-1}$$

Proof. We will prove this bound by building on the analysis of [Li11]. The surface area of a d -dimensional hypersphere is well known to be

$$\mathcal{A}_d = \frac{2\pi^{d/2}}{\Gamma(d/2)},$$

where the gamma function Γ represents the standard extension of the factorial function.

We begin with the following observation of Li:

$$\mathcal{A}_d(\phi) = \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_0^\phi \sin^{d-2} x \, dx.$$

Next, we apply the fact that $\sin x \leq x$ for $x \geq 0$.

$$\begin{aligned} \mathcal{A}_d(\phi) &\leq \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \int_0^\phi x^{d-2} \, dx \\ &= \frac{2\pi^{(d-1)/2}}{\Gamma((d-1)/2)} \left[\frac{1}{d-1} \phi^{d-1} \right] \\ &= \frac{\pi^{(d-1)/2}}{\Gamma((d+1)/2)} \phi^{d-1} \frac{\mathcal{A}_d(\phi)}{\mathcal{A}_d} \\ &\leq \frac{\Gamma(d/2)}{2\sqrt{\pi}\Gamma((d+1)/2)} \phi^{d-1} \end{aligned}$$

We are almost done; we just need to show that $\frac{\Gamma(d/2)}{2\sqrt{\pi}\Gamma((d+1)/2)} \leq 1$. After $\Gamma(3/2)$, subsequent half-values of Γ are increasing and so the desired claim is trivially true for $d \geq 3$. We manually check $d = 1$ and $d = 2$, noting that $\Gamma(1/2) = \sqrt{\pi}$, $\Gamma(1) = 1$, and $\Gamma(3/2) = \sqrt{\pi}/2$.

$$\begin{aligned} \frac{\Gamma(1/2)}{2\sqrt{\pi}\Gamma(2/2)} &= \frac{1}{2} \\ \frac{\Gamma(2/2)}{2\sqrt{\pi}\Gamma(3/2)} &= \frac{1}{\pi} \end{aligned}$$

This completes the proof. □

We are now ready to give the full proof for Theorem 4, which is restated below for convenience.

Theorem 4. *Any algorithm for learning general convex regions incurs a total loss of at least $\Omega(T^{(d-4)/(d-2)})$, even when there are only two regions.*

Proof. Our counterexample depends on two parameters: the dimension d and an error parameter $\epsilon > 0$. We will choose $\epsilon = 1/T^{1/(d-2)}$, where T is the total number of time steps.

Our construction restricts itself to choosing points on \mathcal{S}_d . The construction begins with the opposite points $x_1 = e_1 \triangleq (1, 0, 0, \dots, 0)$ and $x_2 = -e_1$. x_1 has true label 1 and x_2 has true label 2. We will use A to denote the conic hull of the points with true label 1 and B to denote the conic hull of the points with true label 2. By construction, A and B are 0-cones. We will maintain the two invariants that (i) $A \cap B = \{0\}$ and (ii) $A' = B' = \{0\}$, which we will use when invoking Theorem 33.

We now explain how to generate subsequent points x_t for $t \geq 3$. By Theorem 33, we know there is a separating hyperplane \mathcal{H} that passes through the origin and strictly separates $A \setminus \{0\}$ from $B \setminus \{0\}$. We pick an arbitrary hyperplane \mathcal{H} that satisfies the previous statement and examine its intersection with the hypersphere, $X \triangleq \{x \mid \mathcal{H}(x) = 0\} \cap \mathcal{S}_d$.

Since \mathcal{H} passes through the origin, X is a $(d-1)$ -dimensional unit hypersphere. We want to choose our next point x_t to be a point in X that is at least ϵ -far from all previously chosen points x_1, x_2, \dots, x_{t-1} . This condition rules out $(t-1)$ hypersphere caps of angles at most $2 \arcsin \frac{\epsilon}{2}$. Observe that as T increases and ϵ decreases, this angle bound scales as $O(\epsilon)$. By Lemma F.1, we have (hyper-)surface area remaining as long as $(t-1) \cdot O(\epsilon^{d-2}) \leq 1$, i.e. we can pick up to $(1/\epsilon)^{d-2} = T$ total points safely. We will pick an arbitrary point x_t satisfying our ϵ -far rule.

We assign this point x_t a uniform random true label between 1 and 2. It remains to establish that our two invariants are still satisfied. For the sake of contradiction, assume that (i) is no longer true and that there is a non-origin point z in the intersection $A \cap B$. Without loss of generality, x_t was assigned true label 1 and f used to be strictly positive (negative respectively) on $A \setminus \{0\}$ ($B \setminus \{0\}$ respectively). We can deduce the following.

$$\begin{aligned} \mathcal{H}(z) &< 0 \\ \mathcal{H}(z) &= \mathcal{H}\left(\sum_{x \text{ has true label 1}} \alpha_x x\right) \\ &= \sum_{x \text{ has true label 1}} \alpha_x \mathcal{H}(x) \\ &\geq 0 \end{aligned}$$

for some values $\alpha_x \geq 0$. The first inequality follows from the fact that z is in $B \setminus \{0\}$ and the second inequality follows from the fact that all points with true label 1 either are in $A \setminus \{0\}$ before this round or are x_t . We've reached a contradiction and conclude that invariant (i) remains true.

We establish invariant (ii) much in the same way. Again, for the sake of contradiction we assume that (ii) is false. Without loss of generality, assume that x_t was assigned true label 1 and that now there exists a non-origin point z in $A' = A \cap -A$. We follow a similar line of deduction.

$$\begin{aligned} \mathcal{H}(z) &= \mathcal{H}\left(\sum_{x \text{ has true label 1}} \alpha_x x\right) \\ \mathcal{H}(z) &\geq 0 \\ \mathcal{H}(z) &= \mathcal{H}\left(\sum_{x \text{ has true label 1}} \beta_x x\right) \\ \mathcal{H}(z) &\leq 0 \\ \mathcal{H}(z) &= 0 \end{aligned}$$

for some values $\alpha_x \geq 0$ and $\beta_x \leq 0$. But this implies that $\alpha_x = \beta_x = 0$ for all x that have true label 1 and are not the most recent point x_t . We are left with $z = \alpha_{x_t} x_t = \beta_{x_t} x_t$ for some $\alpha_{x_t} \geq 0$ and $\beta_{x_t} \leq 0$, but this can only hold if $\alpha_{x_t} = \beta_{x_t} = 0$ (recall that x_t is not the origin by construction). But then z is the origin, which contradicts our assumption. Hence our assumption is false, i.e. invariant (ii) is indeed maintained.

All that remains is to add up error of our algorithm. Since we chose the true label for the final $T - 2$ points uniformly at random, any online algorithm will make $\frac{T-2}{2}$ mistakes in expectation.

How much error does the algorithm accrue for every mistake? Recall that our construction kept all points at least ϵ away from each other and all points were on the hypersphere (the surface of the unit ball). Figure 1 illustrates the situation with respect to one point. We can solve for the distance that our point of interest x is away from the unit ball minus its cap, which is an upper bound for the convex hull of all other points.

$$\begin{aligned} h(2 - h) &= w^2 \\ 2h &= w^2 + h^2 \\ 2h &= \epsilon^2 \\ h &= \epsilon^2/2 \end{aligned}$$

Hence the algorithm accrues $\Omega(\epsilon^2)$ expected error in each timestep, and so overall any online algorithm is expected to accrue $\Omega(T\epsilon^2)$ total error. By our choice of ϵ , this equals the desired bound. \square

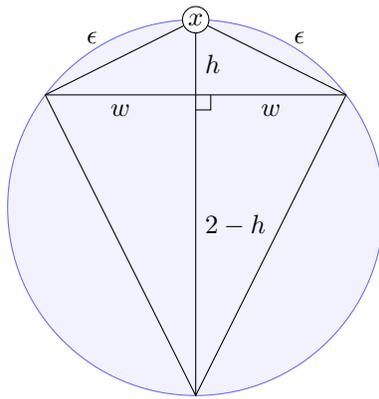


Figure 1: Circular cross-section of the d -dimensional unit hypersphere, \mathcal{S}_d . The two points ϵ away from x (in Euclidean distance) and the diametrically opposite point are marked, forming a kite. The diagonals of this kite result in a right triangle of interest to our analysis, with height h , width w , and hypotenuse ϵ .