JMLR: Workshop and Conference Proceedings vol (2012) 1–15

New Bounds for Learning Intervals with Implications for Semi-Supervised Learning

David P. Helmbold DPH@SOE.UCSC.EDU
Department of Computer Science, University of California at Santa Cruz

Philip M. Long PLONG@SV.NEC-LABS.COM
NEC Labs America

Abstract

We study learning of initial intervals in the prediction model. We show that for each distribution D over the domain, there is an algorithm $\mathcal{A}_D$, whose probability of a mistake in round m is at most $\left(\frac{1}{2}+o(1)\right)\frac{1}{m}$. We also show that the best possible bound that can be achieved in the case in which the same algorithm $\mathcal{A}$ must be applied for all distributions D is at least $\left(\frac{1}{\sqrt{e}}-o(1)\right)\frac{1}{m} > \left(\frac{3}{5}-o(1)\right)\frac{1}{m}$. Informally, "knowing" the distribution D enables an algorithm to reduce its error rate by a constant factor strictly greater than 1. As advocated by Ben-David et al. (2008), knowledge of D can be viewed as an idealized proxy for a large number of unlabeled examples.

Keywords: Prediction model, initial intervals, semi-supervised learning, error bounds.

1. Introduction

Where to place a decision boundary between a cloud of negative examples and a cloud of positive examples is a core and fundamental issue in machine learning. Learning theory provides some guidance on this question, but gaps in our knowledge persist even in the most basic and idealized formalizations of this problem.

Arguably the most basic such formalization is the learning of initial intervals in the prediction model (Haussler et al., 1994). Each concept in the class is described by a threshold θ, and an instance $x \in \Re$ is labeled $+$ if $x \leq \theta$ and $-$ otherwise. The learning algorithm $\mathcal{A}$ is given a labeled m-sample $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ where y_i is the label of x_i. The algorithm must then predict a label $\hat{y}$ for a test point x. The $x_1, ..., x_m$ and x are drawn independently at random from an arbitrary, unknown probability distribution D. Let $\mathrm{opt}(m)$ be the optimal error probability guarantee using m examples in this model. The best previously known bounds (Haussler et al., 1994) were

$$\left(\frac{1}{2}-o(1)\right)\frac{1}{m} \leq \mathrm{opt}(m) \leq (1+o(1))\,\frac{1}{m}. \tag{1}$$

To our knowledge, this factor of 2 gap has persisted for nearly two decades.

The proof of the lower bound of (1) uses a specific choice of D, no matter what the target. Thus, it also lower bounds the best possible error probability guarantee for algorithms that are given the distribution D as well as the sample. The upper bound holds for a particular algorithm and all distributions D, so it is also an upper bound on the best possible error probability guarantee when the algorithm does not know D.

The increasing availability of unlabeled data has inspired much recent research on the question of how to use such data, and the limits on its usefulness. Ben-David et al. (2008) proposed knowledge of the distribution D as a clean, if idealized, proxy for access to large numbers of unlabeled examples. Since the lower bound in (1) did not exploit the algorithm's lack of a priori knowledge of D, improved lower bounds exploiting this lack of knowledge may also shed light on the utility of unlabeled data.

This paper studies how knowledge of D affects the error rate when learning initial intervals in the prediction model. Our positive result is an algorithm that, *when given* D, has an error rate at most $\left(\frac{1}{2} + o(1)\right)\frac{1}{m}$. This matches the lower bound of (1) up to lower order terms. As a complementary negative result, we show that *any* prediction algorithm *without prior knowledge of* D can be forced to have an error probability at least $\left(\frac{1}{\sqrt{e}} - o(1)\right)\frac{1}{m} > \left(\frac{3}{5} - o(1)\right)\frac{1}{m}$. Thus not knowing D leads to *at least* a 20% increase in the probability of making a mistake (when the target and distribution are chosen adversarially). A third result shows that the maximum margin algorithm can be forced to have the even higher error rate $(1 - o(1))\frac{1}{m}$.

The training data reduces the version space (the region of potential values of θ) to an interval between the greatest positive example and the least negative example. Furthermore, all examples outside this region are classified correctly by any θ in the version space. Our algorithm achieving the $\left(\frac{1}{2} + o(1)\right)\frac{1}{m}$ error probability protects against the worst case by choosing a hypothesis $\hat{\theta}$ in the middle (with respect to the distribution D) of this region of uncertainty. This can be viewed as a D-weighted halving algorithm and follows the general principle of getting in the "middle" of the version space (Herbrich et al., 2001; Kääriäinen, 2005).

As has become common since (Ehrenfeucht et al., 1989), the proof of our $\left(\frac{1}{\sqrt{e}} - o(1)\right)\frac{1}{m}$ lower bound proceeds by choosing θ and D randomly and analyzing the error rate of the resulting Bayes optimal algorithm. The distribution D in our construction concentrates a moderate amount q of probability very close to one side or the other of the decision boundary. If D is unknown, and no examples are seen from the accumulation point (likely if q is not too large), the algorithm cannot reliably "get into the middle" of the version space. Unlike most analyses showing the benefits of semi-supervised learning that rely on an assumption of sparsity near the decision boundary, our analysis uses distributions that are peaked at the decision boundary.

Related work. Learning from a labeled sample and additional unlabeled examples is called semi-supervised learning. Semi-supervised learning is an active and diverse research area; see standard texts like (Chapelle et al., 2006) and (Zhu and Goldberg, 2009) for more information.

Our work builds most directly on the work of Ben-David et al. (2008). They proposed using knowledge of D as a proxy for access to a very large number of unlabeled examples, and considered how this affects the sample complexity of learning in the PAC model, which is closely related to the error rate in the prediction model (Haussler et al., 1994). Their main results concerned limitations on the impact of the knowledge of D; Darnstädt and Simon (2011) extended this line of research, also demonstrating such limitations. In contrast, the thrust of our main result is the opposite, that knowledge of D gives at least a 16% reduction in the (worst-case) prediction error rate. Balcan and Blum (2010) introduced a framework

to analyze cases in which partial knowledge of relationship between the distribution and the target classifier can improve error rate, whereas the focus of this paper is to study the benefits of knowing D even potentially in the absence of a relationship between D and the target. Kääriäinen (2005) analyzed algorithms that use unlabeled data to estimate the metric $\rho_D(f,g) \stackrel{\text{def}}{=} \Pr_{x\sim D}(f(x) \neq g(x))$, and then choose a hypothesis at the "center" of the version space of classifiers agreeing with the labels on the training examples. Our algorithm for learning given knowledge of D follows this philosophy. Urner et al. (2011) proposed a framework to analyze cases in which unlabeled data can help to train a classifier that can be evaluated more efficiently.

The upper bound of (1) is a consequence of a more general bound in terms of the VC-dimension. Li et al. (2001) showed that the leading constant in the general bound cannot be improved, even in the case that the VC-dimension is 1. However their construction uses tree-structured classes that are more complicated than the initial intervals studied here.

2. Further Preliminaries and Main Results

For any particular D and θ, the expected error of Algorithm A, $\mathrm{Err}_m(\mathcal{A}; D, \theta)$, is the probability that its prediction is not the correct label of x, where the probability is over the $m+1$ random draws from D and any randomization performed by $\mathcal{A}$. We are interested in the worst-case error of the best algorithm: if the algorithm can depend on D, this is

$$\mathrm{opt}_D(m) = \inf_{\mathcal{A}} \sup_{\theta} \mathrm{Err}_m(\mathcal{A}; D, \theta),$$

and, if not,

$$\mathrm{opt}(m) = \inf_{\mathcal{A}} \sup_{D,\theta} \mathrm{Err}_m(\mathcal{A}; D, \theta).$$

Our main lower bound is the following.

Theorem 1 $\mathrm{opt}(m) \geq \left(\frac{1}{\sqrt{e}} - o(1)\right)\frac{1}{m} \geq \left(\frac{3}{5} - o(1)\right)\frac{1}{m}$.

This means that for every algorithm learning initial intervals, there is a distribution D and threshold θ such that the algorithm's mistake probability (after seeing m examples, but not the distribution D) is at least $\left(\frac{1}{\sqrt{e}} - o(1)\right)\frac{1}{m}$. In our proof, distribution D depends on m.

Our main upper bound is the following.

Theorem 2 *For all probability distributions D,* $\mathrm{opt}_D(m) \leq \left(\frac{1}{2} + o(1)\right)\frac{1}{m}$.

We show Theorem 2 by analyzing an algorithm that gets into the middle of the version space with respect to the given distribution D.

When there are both positive and negative examples, a *maximum margin algorithm* (Vapnik and Lerner, 1963; Boser et al., 1992) makes its prediction using a hypothesized threshold $\hat{\theta}$ that is halfway between the greatest positive example, and the least negative example.

Theorem 3 *For any maximum margin algorithm A, there is a D and a θ such that*

$$Err_m(\mathcal{A}_{\mathrm{MM}}; D, \theta) \geq (1 - o(1))\,\frac{1}{m}.$$

Our construction chooses D and θ as functions of m. Note that in the case of intervals, the maximum margin algorithm is similar to the (un-weighted) halving algorithm.

Throughout we use U to denote the uniform distribution on the open interval $(0, 1)$.

3. Proof of Theorem 1

As mentioned in the introduction, we will choose the target θ and the distribution D randomly, and prove a lower bound on the Bayes optimal algorithm when D and θ are chosen in this way. No algorithm A can do better on average over the random choice of θ and D than the Bayes optimal algorithm that knows the *distributions* over D and θ. This in turn implies that for any algorithm A there exist a particular θ and D for which the lower bound holds for A.

Let $q = c/m$ for a $c \in [0, m)$ to be chosen later. Define distribution $D_{\theta,q}(x)$ as the following mixture:

- with probability $p = 1 - q$, x is drawn from U, the uniform distribution on $(0, 1)$.
- with probability $q = c/m$, $x = \theta$.

We will analyze the following:

1. Fix the sample size m.
2. Draw θ from U and set the target to be $(-\infty, \theta]$ with probability $1/2$, and $(-\infty, \theta)$ with probability $1/2$.
3. Draw an m-sample $S = \{x_1, \ldots, x_m\}$ from $D^m_{\theta,q}$. Extend S to the extended sample, S^+ by adding $x_0 = 0$ and $x_{m+1} = 1$ to S. The *labeled sample* $\mathcal{L} = \{(x_i, y_i)\}$ where $x_i \in S^+$ and y_i is the label of x_i given by the target (either $(-\infty, \theta)$ or $(-\infty, \theta]$).
4. Draw a final test point x also iid from $D_{\theta,q}$.

This setting, which we call the *open-closed experiment*, is not "legal", because it sometimes uses open intervals as targets. However, we will now show that a lower bound for this setting implies a similar lower bound when only closed initial intervals are used. We define the *legal experiment* as above, except that instead of using the open target $(-\infty, \theta)$, the adversary uses target $(-\infty, \theta - \frac{1}{m^3}]$.

Lemma 4 *For any algorithm A and any number m of examples, let p_{legal} be the probability that A makes a mistake in the legal experiment, and p_{oc} be the probability that A makes a mistake in the open-closed experiment. Then $p_{\text{legal}} \geq p_{\text{oc}} - \frac{m+1}{m^3}$.*

Proof. If none of the training or test examples falls (strictly) between $\theta - \frac{1}{m^3}$ and θ, then the training and test data are the same in both experiments. The probability that this happens is at least $1 - \frac{m+1}{m^3}$. ∎

Since a $(C - o(1))/m$ lower bound for the open-closed experiment implies such a bound for the legal experiment, we can concentrate on the open-closed experiment.

We now consider the following events.

MISTAKE	is the event that the Bayes Optimal classifier predicts incorrectly.
ZERO	is the event that no $x_i \in S$ equals θ.
ONE	is the event that exactly one $x_i \in S$ equals θ.

Now, using these events,

$$\Pr(\textsc{mistake}) \geq \Pr(\textsc{zero}) \Pr(\textsc{mistake} \mid \textsc{zero}) + \Pr(\textsc{one}) \Pr(\textsc{mistake} \mid \textsc{one}). \quad (2)$$

3.1. Event ZERO, no $x_i \in S$ is equal to θ.

First we lower bound the probability of ZERO.

Lemma 5 $\Pr(\textsc{zero}) = (1 - c/m)^m \geq e^{-c}\left(1 - \frac{c^2}{m}\right)$.

Proof: The draws from $D_{\theta,q}$ are independent so for $q = c/m$ we have $\Pr(\textsc{zero}) = p^m = (1 - c/m)^m$. Furthermore, $\ln\left((1 - c/m)^m\right) = m \ln(1 - c/m) \geq m\left(-\frac{c}{m} - \frac{c^2}{m^2}\right) = -c - \frac{c^2}{m}$ and $\exp(-c^2/m) \geq 1 - c^2/m$. ∎

Now, we lower bound the probability of a mistake, given ZERO. In this case, the x_i's in S and θ are all drawn i.i.d. from the uniform distribution on $(0, 1)$, and with probability one the x_i are all different from θ and each other. Let x_+ be the largest $x_i \in S^+$ smaller than θ, and let x_- be smallest $x_i \in S^+$ greater than θ. (Recall that instances 0 and 1 have been added to S^+ so x_+ and x_- are well-defined.)

Lemma 6 *Assume event* ZERO *and let x_+ be largest positive point in S^+ and x_- be the smallest negative point in S^+. After conditioning on the labeled sample $\mathcal{L}$, if the test point is sampled from $D_{\theta,q}$ (independent of S^+) then the error probability of the Bayes Optimal predictor is*

$$\Pr(\textsc{mistake} \mid x_+, x_-) = \frac{q}{2} + p\left(\frac{x_- - x_+}{4}\right) = \frac{c}{2m} + \frac{(m-c)(x_- - x_+)}{4m}.$$

Proof: After conditioning on $\mathcal{L}$ and event ZERO, θ is uniformly distributed on (x_+, x_-).

The label of test point x is known whenever $x \leq x_+$ or $x \geq x_-$. Only when $x_+ < x < x_-$ can the Bayes optimal predictor make a mistake.

The Bayes optimal predictor predicts + on x if $x < (x_+ + x_-)/2$, predicts − on x if $x > (x_+ + x_-)/2$, and predicts arbitrarily when $x = (x_+ + x_-)/2$. Therefore, for a given value of θ, when the test point x is drawn from $D_{\theta,q}$, the probability of mistake is $q/2$ (for the fraction of time that $x = \theta$) plus $p \cdot \left|\frac{x_+ + x_-}{2} - \theta\right|$ for the chance that x is drawn from U and falls between θ and the midpoint of (x_+, x_-).

$$\begin{aligned}
\Pr(\textsc{mistake} \mid x_+, x_-) &= \int_{x_+}^{\frac{x_+ + x_-}{2}} \left(\frac{q}{2} + p\left(\frac{x_+ + x_-}{2} - \theta\right)\right) dP(\theta \mid \theta \in [x_+, x_-]) \\
&\quad + \int_{\frac{x_+ + x_-}{2}}^{x_-} \left(\frac{q}{2} + p\left(\theta - \frac{x_+ + x_-}{2}\right)\right) dP(\theta \mid \theta \in [x_+, x_-]) \\
&= \frac{q}{2} + \frac{p\,(x_- - x_+)}{4}
\end{aligned}$$

as desired. ∎

The following lemma is due to Moran (1947).

Lemma 7 (Moran, 1947) *Assume a set $\{x_1, \ldots, x_m\}$ of $m \geq 3$ points are drawn iid from the uniform distribution on the unit interval. Relabel these points so that $x_1 \leq x_2 \leq \cdots \leq x_m$ and set $x_0 = 0$ and x_{m+1}=1. For the $m+1$ gap lengths defined by $g_i = x_{i+1} - x_i$ for $0 \leq i \leq m$, we have $\mathbb{E}\left(\sum_{i=0}^{m} g_i^2\right) = \frac{2}{m+2}$.*

Note: if the set of points is drawn from a circle, then the first point can be taken as the "endpoint", splitting the circle into an interval. This leads to the following, which we will find useful later.

Lemma 8 *The expected sum of squared arc-lengths when the circle of circumference 1 is partitioned by $m \geq 4$ random points from the uniform distribution is $2/(m+1)$.*

Coming back to our lower bound proof, we are now ready to prove a lower bound for the ZERO case.

Lemma 9 *For the Bayes Optimal Predictor,*

$$\Pr(\text{MISTAKE} \mid \text{ZERO}) = \frac{c}{2m} + \frac{1}{2(m+2)} - \frac{c}{2m(m+2)}.$$

Proof: Given the event ZERO, each x_i in S is drawn iid from the uniform distribution on $(0,1)$. We find it convenient to relabel the $x_i \in S$ in sorted order so that $x_1 \leq x_2 \cdots \leq x_m$. Note that $x_0 = 0$ and $x_{m+1} = 1$ in S^+ are defined consistently with this sorted order. To simplify the notation, we leave the conditioning on event ZERO implicit in the remainder of the proof.

$$\Pr(\text{MISTAKE}) = \int\int \Pr(\text{MISTAKE} \mid S, \theta)\, d\theta\, dP(S) \tag{3}$$

$$= \int \sum_{i=0}^{m} \Pr(\theta \in (x_i, x_{i+1}) \mid S) \Pr(\text{MISTAKE} \mid \theta \in (x_i, x_{i+1}), S)\, dP(S) \tag{4}$$

Since the threshold θ is drawn from U on $(0,1)$, $\Pr(\theta \in (x_i, x_{i+1}) \mid S) = x_{i+1} - x_i$. With an application of Lemma 6 we get:

$$\sum_{i=0}^{m} \Pr(\theta \in (x_i, x_{i+1}) \mid S) \Pr(\text{MISTAKE} \mid \theta \in (x_i, x_{i+1}), S)$$

$$= \sum_{i=0}^{m} (x_{i+1} - x_i) \left(\frac{q}{2} + \frac{p(x_{i+1} - x_i)}{4}\right) = \frac{q}{2} + \frac{p}{4} \sum_{i=0}^{m} (x_{i+1} - x_i)^2.$$

Substituting this into (4),

$$\Pr(\text{MISTAKE}) = \int \left(\frac{q}{2} + \frac{p}{4} \sum_{i=0}^{m} (x_{i+1} - x_i)^2\right) dP(S) \tag{5}$$

$$= \frac{q}{2} + \frac{p}{4} \cdot \mathbb{E}_{S \sim U^m} \left[\sum_{i=0}^{m} (x_{i+1} - x_i)^2\right] \tag{6}$$

$$= \frac{q}{2} + \frac{p}{2(m+2)} \tag{7}$$

using Lemma 7 to evaluate the expectation in (6). Replacing q by c/m and p by $1 - c/m$ gives the desired result. ∎

3.2. Event ONE, exactly one $x_i \in S$ is equal to θ.

We lower bound the second term on the RHS of (2) by bounding $\Pr(\text{ONE})$ and $\Pr(\text{MISTAKE} \mid \text{ONE})$.

Lemma 10 $\Pr(\text{ONE}) = mq(1-q)^{m-1} \geq ce^{-c}\left(1 + \frac{c-c^2}{m} - \frac{c^3}{m^2}\right)$.

Proof:

$$\Pr(\text{ONE}) = mq(1-q)^{m-1} = \frac{c(1-c/m)^m}{1-c/m} \geq \frac{ce^{-c}(1-c^2/m)}{1-c/m} \geq ce^{-c}\left(1 + \frac{c-c^2}{m} - \frac{c^3}{m^2}\right),$$

where Lemma 5 was used to bound $(1-c/m)^m$ and the last inequality used the fact that $1/(1-z) \geq 1+z$ for all $z \in [0,1)$. ∎

Lemma 11 $\Pr(\text{MISTAKE} \mid \text{ONE}) \geq \frac{1}{2m} \cdot \frac{m-1}{m+1} = \frac{1}{2m} - O(1/m^2)$.

Proof: Since θ is drawn from the uniform distribution, after conditioning on the event ONE, the x_i in S are i.i.d. from the uniform distribution on $[0,1]$. Again consider the points relabeled in ascending order. Each of the $x_i \in S$ is equally likely to be θ, and the $x_i = \theta$ is equally likely to be labeled $+$ or $-$, giving $2m$ equally likely possibilities. As before, define x_+ and x_- to be the largest $x_i \in S^+$ labeled $+$ and the smallest $x_i \in S^+$ labeled $-$ respectively. We proceed assuming the Bayes Optimal Algorithm "knows" that one of the $x_i = \theta$, i.e. that event ONE occurred. (This can only reduce its error probability.) To simplify the notation, we leave the conditioning on event ONE implicit in the remainder of the proof.

If there are both positive and negative examples, so that there is a greatest positive example x_+ and a least negative examples x_-, there are two possibilities: either the target is $(-\infty, x_+]$ or it is $(-\infty, x_-)$. In either case, the entire open interval between x_+ and x_- shares the same label, and since the two cases are equally likely that label is equally likely to either $+$ or $-$. Therefore:

$$\Pr(\text{MISTAKE}) \geq \int \sum_{i=1}^{m-1} \Pr(x_i = x_+ \mid S) \Pr(\text{MISTAKE} \mid S, x_+ = x_i)\, d\Pr(S) \tag{8}$$

$$= \int \sum_{i=1}^{m-1} \Pr(x_i = x_+ \mid S) \left(\frac{x_{i+1} - x_i}{2}\right) d\Pr(S). \tag{9}$$

We consider the sum in more detail. Note that x_i can be x_+ when either $x_i = \theta$ and is labeled $+$ or $x_{i+1} = \theta$ and is labeled $-$.

$$\sum_{i=1}^{m-1} \Pr(x_i = x_+ \mid S) \left(\frac{x_{i+1} - x_i}{2}\right) = \sum_{i=1}^{m-1} \frac{1}{m} \left(\frac{x_{i+1} - x_i}{2}\right) = \frac{1}{2m}(x_m - x_1).$$

Plugging it into (9) gives: $\Pr(\text{MISTAKE}) = \frac{1}{2m}\mathbb{E}_{S\sim U^m}[x_m - x_1] = \frac{1}{2m} \cdot \frac{m-1}{m+1}$ since the expected length of each missing end-interval is $1/(m+1)$. ∎

3.3. Putting it together

Combining Lemma 5, Lemma 9, Lemma 10, and Lemma 11 we get that $\Pr(\text{MISTAKE} \mid \text{ZERO})\Pr(\text{ZERO}) + \Pr(\text{MISTAKE} \mid \text{ONE})\Pr(\text{ONE})$ is at least

$$\left(\frac{c}{2m} + \frac{1}{2(m+2)} - \frac{c}{2m(m+2)}\right) e^{-c}\left(1 - \frac{c^2}{m}\right) + \frac{m-1}{2m(m+1)} c e^{-c}\left(1 + \frac{c-c^2}{m} - \frac{c^3}{m^2}\right)$$
$$= \frac{1+2c}{2m} e^{-c} - O(1/m^2).$$

Setting $c = 1/2$, the maximizer of $(1+2c)e^{-c}$, the bound becomes $1/(m\sqrt{e}) - O(1/m^2)$, completing the proof of Theorem 1.

4. Proof of Theorem 2

Here we show that knowledge of D can be exploited by a maximum-margin-in-probability algorithm to achieve prediction error probability at most $(1/2 + o(1))/m$. The marginal distribution D and target threshold θ are chosen adversarially, but the algorithm is given distribution D as well as the training sample.

The first step is to show that we can assume without loss of generality that D is the uniform distribution U over $(0,1)$. This is a slight generalization of the "rescaling trick" of Ben-David et al. (2008).

Lemma 12 $\text{opt}_D(m) \leq \text{opt}_U(m)$.

Proof Our proof is through a prediction-preserving reduction (Pitt and Warmuth, 1990). This consists of a (possibly randomized) instance transformation ϕ and a target transformation ψ. In this proof ϕ maps $\mathbf{R}$ into $[0,1]$, and ψ maps a threshold in $\mathbf{R}$ to a new threshold in $[0,1]$. Implicit in the analysis of Pitt and Warmuth (1990) is the observation that, if $\phi(x) \leq \psi(\theta) \Leftrightarrow x \leq \theta$ for all training and test examples x, then an algorithm A_t with prediction error bound b_t for the transformed problem can be used to solve the original problem. By feeding A_t the training data $\phi(x_1), ..., \phi(x_m)$ and the test point $\phi(x)$, and using A_t's prediction (of whether $\phi(x) \leq \psi(\theta)$ or not) one gets an algorithm with prediction error bound b_t for the original problem.

When D has a density, $\phi(x) = \Pr_{z\sim D}(z \leq x)$ and $\psi(\theta) = \Pr_{z\sim D}(z \leq \theta)$. In this case the distribution $\phi(x)$ is uniform over $(0,1)$, and since ϕ and ψ are identical and monotone, $\phi(x) \leq \psi(\theta) \Leftrightarrow x \leq \theta$.

If D has a one or more accumulation points, then, for each accumulation point x_a, we choose $\phi(x)$ uniformly from the interval $(\Pr_{z\sim D}(z < x_a), \Pr_{z\sim D}(z \leq x_a)]$. We still set $\psi(\theta) = \Pr_{z\sim D}(z \leq \theta)$ everywhere. For this transformation, the probability distribution over $\phi(x)$ is still uniform over $(0,1)$, and, as before, if $x \neq \theta$, or if x is not an accumulation point, then $\phi(x) \leq \psi(\theta) \Leftrightarrow x \leq \theta$. When $x_a = \theta$ for an accumulation point x_a, $\phi(x_a) \leq \psi(\theta)$, since $\psi(\theta)$ is set to be the right endpoint of $(\Pr_{z\sim D}(z < x_a), \Pr_{z\sim D}(z \leq x_a)]$. Thus, overall, $\phi(x) \leq \psi(\theta) \Leftrightarrow x \leq \theta$, and the probability of a mistake in the original problem is bounded by the mistake probability with respect to the transformed problem. ■

So now we are faced with the subproblem of learning initial intervals in the case that D is the uniform distribution U over $(0,1)$. The algorithm that we analyze for this problem is the following maximum margin algorithm $\mathcal{A}_{\text{MM}}$: if the training data includes both positive and negative examples, it predicts using a threshold halfway between the greatest positive example and the least negative example. If all of the examples are negative, it uses a threshold halfway between the least negative example and 0, and if all of the examples are positive, $\mathcal{A}_{\text{MM}}$ uses a threshold halfway between the greatest positive example and 1.

The basic idea exploits the fact that the uniform distribution is invariant to horizontal shifts to average over random shifts. We define $x \oplus s \equiv x + s - \lfloor x + s \rfloor$ to be addition modulo 1, so that, intuitively, $x \oplus s$ is obtained by starting at x, and walking s units to the right while wrapping around to 0 whenever 1 is reached. We extend the $\oplus$ notation to sets in the natural way: if $T \subseteq [0,1)$ then $T \oplus s = \{t \oplus s : t \in T\}$.

Fix an arbitrary target threshold $\theta \in (0,1)$ and also fix (for now) an arbitrary set $S = \{x_1, \ldots, x_m\}$ of m training points (whose labels are determined by θ). Renumber the points in S so that $x_1 \leq x_2 \leq \cdots \leq x_m$. To simplify some expressions, we use both x_1 and x_{m+1} to refer to x_1. For each $x, s \in [0,1]$, let $\mathit{error}(s,x)$ be the $\{0,1\}$-valued indicator for whether $\mathcal{A}_{\text{MM}}$ makes a mistake when trained with $S \oplus s$ and tested with x. Let $\mathit{error}(s) = \mathbb{E}_{x \sim U}(\mathit{error}(s,x))$, and let $\mathit{error} = \mathbb{E}_{s \sim U}(\mathit{error}(s))$.

For $1 \leq i < m$, let $G_i = [x_i, x_{i+1})$ be the points in the interval between x_i and x_{i+1}, and let $G_m = [x_m, 1) \cup [0, x_1)$, so the G_i partition $[0,1)$. Let $R_i = \{s \in [0,1) : \theta \in G_i \oplus s\}$ and notice that the R_i also partition $[0,1)$. We have $\mathit{error} = \int_0^1 \mathit{error}(s)ds = \sum_i \int_{s \in R_i} \mathit{error}(s)ds$.

We now consider $\int_{s \in R_i} \mathit{error}(s)ds$ in more detail. This integral corresponds to the situation where the shifted sample causes θ to fall in the "gap" between x_i and x_{i+1}. Depending on the location of θ and the length of the gap, the shifted interval might extend past either 0 or 1 while containing θ, and thus "wrap around" to the other side of the unit interval.

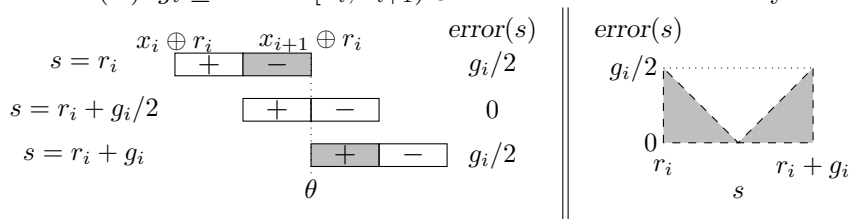
Let the gap length g_i be $x_{i+1} - x_i$ (or $1 - x_m + x_1$ if $i = m$), and let r_i be the shift taking x_{i+1} to θ, so $x_{i+1} \oplus r_i = \theta$. Note that a shift of $r_i + g_i$ takes x_i to θ even though $r_i + g_i$ might be greater than one, which happens when $\theta \in [x_i, x_{i+1})$.

We will now give a "proof-by-plot" that $\int_{s \in R_i} \mathit{error}(s)ds \leq g_i^2/4$. For an algebraic proof, see Appendix A. We begin by assuming $\theta \leq 1/2$ since the situation with $\theta > 1/2$ is symmetrical. The cases we consider depend on the relationship between g_i and θ: case (A) $g_i \leq \theta$, case (B) $\theta < g_i \leq 1 - \theta$, and case (C) $1 - \theta < g_i$. Cases (B) and (C) have two sub-cases depending on whether or not $\theta \leq g_i/2$.

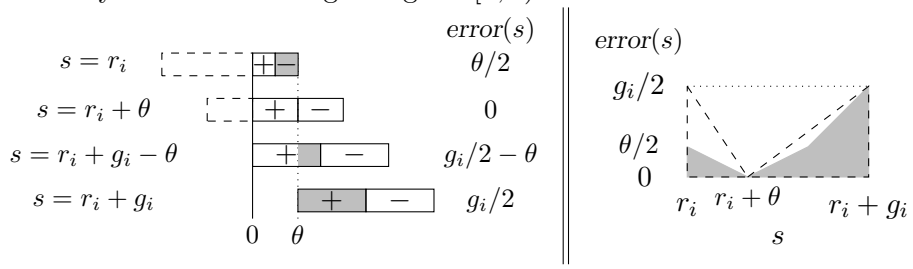
The figure for each case shows several shiftings of the interval, and plots $\mathit{error}(s)$ as a function of s. The predictions of $\mathcal{A}_{\text{MM}}$ on the shifted interval are indicated, and the region where $\mathcal{A}_{\text{MM}}$ makes a prediction error is shaded. The top shifting in each case is actually for $s = r_i$ plus some small ϵ (so that $\theta < x_{i+1} \oplus s$), although it is labeled as $s = r_i$ for simplicity.

Note that in each case, the plot of $\mathit{error}(s)$ lies within two triangles with height $g_i/2$. Therefore the integrals $\int_{r_i}^{r_1 + g_i} \mathit{error}(s)ds$ are at most $g_i^2/4$.

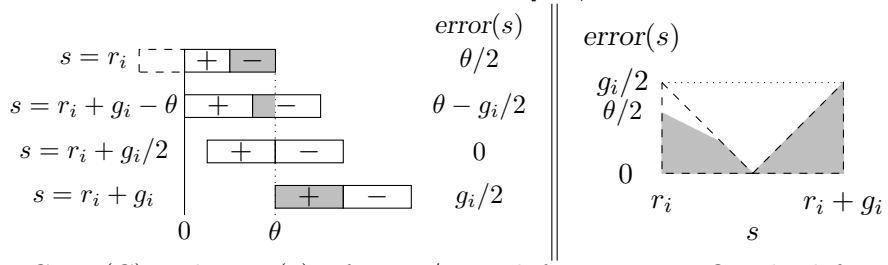
Case (A): $g_i \le \theta$. The $[x_i, x_{i+1}) \oplus s$ interval intersects θ only when $s + x_i \le \theta < s + x_{i+1}$.



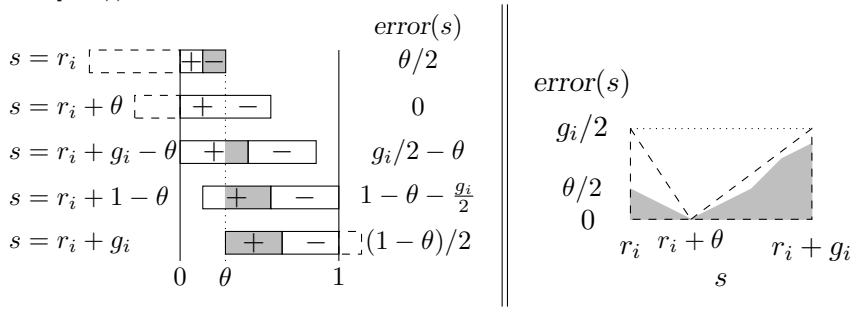
Case (B), subcase (1): $g_i + \theta \le 1$ and $\theta < g_i/2$. Note that the dashed part of the interval is actually shifted to the right-edge of $[0, 1)$.



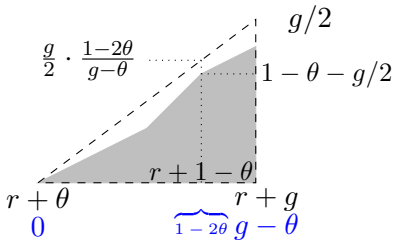
Case (B), subcase (2): $g_i + \theta \le 1$ and $g_i/2 < \theta < g_i$. Again, the dashed part of the interval is actually shifted to right-edge of $[0, 1)$.



Case (C), subcase (1): $\theta < g_i/2$, and $\theta + g_i > 1$. On the left: shifting the interval between x_i and x_{i+1} to intersect θ. In this case not only can part of the interval be shifted to right-edge of $[0, 1)$, but part of the interval can also extend beyond 1 (and be shifted to the left-edge of $[0, 1)$) while the shifted interval contains θ.

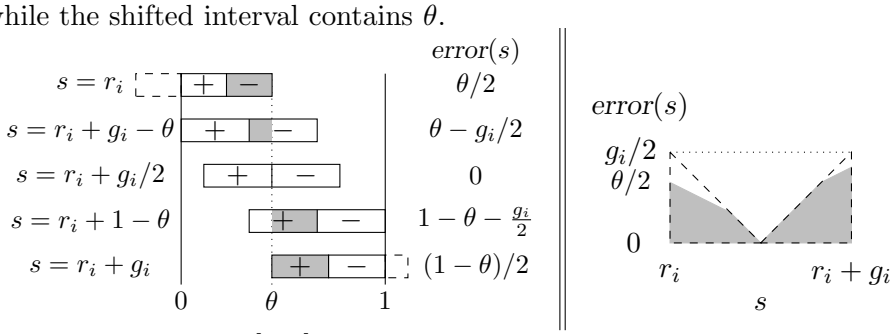


The correctness of the plot relies on the "bump" in the plot at $s = r_i + 1 - \theta$ never rising outside of the triangle. We have redrawn the triangle in more detail to the right (dropping the subscripts), and relabeled the x-axis (in blue). This shows us that at the "bump", the dotted line of the triangle is at height $\frac{g}{2} \cdot \frac{1-2\theta}{g-\theta}$ while the bump is at height $1 - \theta - g/2$.



Solving $\frac{g}{2} \cdot \frac{1-2\theta}{g-\theta} = 1 - \theta - g/2$ for g yields $g = 2\theta$ and $g = 1 - \theta$, exactly the boundaries of this case. Therefore the difference $\frac{g}{2} \cdot \frac{1-2\theta}{g-\theta} - 1 - \theta - g/2$ (which is the amount by which the boundary of the triangle lies above the bump) does not change sign. When $g = 5/6$ and $\theta = 2/6$ the difference is $1/36$, so the difference remains non-negative for all g_i and θ in this case.

Case (C), subcase (2): $\theta > g_i/2$, and $\theta + g_i > 1$. On the left: shifting the interval between x_i and x_{i+1} to intersect θ. Again, parts of the interval can be shifted across 0 and across 1 while the shifted interval contains θ.



We now have, for any $\theta \in [0, 1]$,

$$\begin{aligned} \Pr_{S\sim U^m, x\sim U}(\mathcal{A}_{\mathrm{MM}} \text{ incorrect}) &= \mathbb{E}_{S\sim U^m}\mathbb{E}_{s\sim U, x\sim U}(\mathcal{A}_{\mathrm{MM}}(S \oplus s, x) \neq 1_{\leq\theta}(x)) \\ &\leq \mathbb{E}_{S\sim U^m} \sum_{i=1}^{m} g_i^2/4 \;\leq\; \frac{1}{2m+2} \end{aligned}$$

where the last inequality uses Lemma 8. This completes the proof of Theorem 2.

5. Proof of Theorem 3

In this section we show that any maximum margin algorithm can be forced to have an error probability $(1 - o(1))/m$ when it is not given knowledge of D (*i.e.* without transforming the input as in the previous section). This is a factor of 2 worse than our upper bound for an algorithm that uses knowledge D to maximize a probability-weighted margin.

Let c be a positive even integer (by choosing a large constant value for c, our lower bound will get arbitrarily close to $\frac{1}{m}$). For a given training set size m, consider the set $T = \{3^{-cm}, \ldots, 3^{-2}, 3^{-1}\}$ containing the first cm powers of $1/3$.

Fix distribution D to be the uniform distribution over T. Fix the target threshold to be $\theta = 3^{-cm/2-1}$, so that half the points in T are labeled positively (recall that all points less than or equal to the threshold are labeled positively).

Maximum margin algorithms can make different predictions only when all the examples have the same label. For this choice of a distribution D and a target θ, the probability that all the examples have the same label is 2×2^{-m}. Thus for large enough m, the difference between maximum margin algorithms is negligible. From here on, let us consider the algorithm $\mathcal{A}_{\mathrm{MM}}$, defined earlier, that adds two artificial examples, $(0, +)$ and $(1, -)$, and predicts using the maximum margin classifier on the resulting input.

For $1 \leq i \leq cm/2$, let T_i be the i points in T just above the threshold: if $\ell = -cm/2-1 = \log_3 \theta$ then $T_i = \{3^{\ell+1}, 3^{\ell+2}, \ldots, 3^{\ell+i}\}$. Let event $\textsc{Miss}_i$ be the event that none of the m

training points are in T_i. For $i < cm/2$, let event EXACT_i be the event that both (a) none of the m training points are in T_i, *and* (b) some training point is in T_{i+1} (i.e. some training point is $3^{\ell+i+1}$). Let $\text{EXACT}_{cm/2}$ be the event that no training point is labeled "$-$". Therefore the EXACT_i events are disjoint and $\text{MISS}_i = \bigcup_{j \geq i} \text{EXACT}_j$.

Note that if EXACT_i occurs, then the smallest negative example is $3^{\ell+i+1}$. Furthermore, all points in T_i are less then half this value and the maximum margin algorithm predicts incorrectly on exactly the i points in T_i, so $\Pr(\textit{error}|\text{EXACT}_i) = i/(cm)$. Thus, for $m > 2c$, we have

$$\Pr(\textit{error}) = \sum_{i=1}^{cm/2} \Pr(\textit{error}|\text{EXACT}_i) \Pr(\text{EXACT}_i) = \sum_{i=1}^{cm/2} \frac{i}{cm} \Pr(\text{EXACT}_i)$$
$$= \frac{1}{cm} \sum_{i=1}^{cm/2} \Pr(\text{MISS}_i) = \frac{1}{cm} \sum_{i=1}^{cm/2} \left(\frac{cm-i}{cm}\right)^m \geq \frac{1}{cm} \sum_{i=1}^{c^2} \left(1 - \frac{i/c}{m}\right)^m .$$

For $i \leq c^2$, in the limit as $m \to \infty$, $\left(1 - \frac{i/c}{m}\right)^m \to \exp(-i/c)$, so for large enough m (large relative to the constant c) we can continue as follows.

$$\begin{aligned}
\Pr(\textit{error}) &\geq \frac{1}{cm} \sum_{i=1}^{c^2} (1-\epsilon) \exp(-i/c) = \frac{1-\epsilon}{cm} \sum_{i=1}^{c^2} \exp(-1/c)^i \\
&= \frac{1-\epsilon}{cm} \exp(-1/c) \frac{1-\exp(-1/c)^{c^2}}{1-\exp(-1/c)} = \frac{1-\epsilon}{cm} \exp(-1/c) \frac{1-\exp(-c)}{1-\exp(-1/c)} \\
&= \frac{(1-\epsilon)(1-\epsilon_2)}{cm} \frac{\exp(-1/c)}{1-\exp(-1/c)}
\end{aligned}$$

where $\epsilon_2 = e^{-c}$. Now, replacing $1/c$ by a we get: $\Pr(\textit{error}) \geq \frac{(1-\epsilon)(1-\epsilon_2)}{m} \frac{a \exp(-a)}{1-\exp(-a)}$ Using L'Hopitals rule, we see that the limit of the second fraction as $a \to 0$ is 1. So for large enough c, the second fraction is at least $1 - \epsilon_3$ and $\Pr(\textit{error}) \geq \frac{(1-\epsilon)(1-\epsilon_2)(1-\epsilon_3)}{m}$. Thus, by making the constant c large enough, and choosing m large enough compared to c, the expected error of the maximum margin algorithm can be made arbitrarily close to $1/m$.

6. Conclusion

Algorithms that know the underlying marginal distribution D over the instances can learn significantly more accurately than algorithms that do not. Since knowledge of D has been proposed as a proxy for a large number of unlabeled examples, our results indicate a benefit for semi-supervised learning. It is particularly intriguing that our analysis shows the benefit of semi-supervised learning when the distribution is nearly uniform, but slightly concentrated near the decision boundary. This is in sharp contrast to previous analyses showing the benefits of semi-supervised learning, which typically rely on a "cluster assumption" postulating that examples are sparse along the decision boundary.

References

M. Balcan and A. Blum. A discriminative model for semi-supervised learning. *JACM*, 57 (3), 2010.

S. Ben-David, T. Lu, and D. Pal. Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. *COLT*, 2008.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. *Proceedings of the 1992 Workshop on Computational Learning Theory*, pages 144–152, 1992.

O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, 2006.

Malte Darnstädt and Hans-Ulrich Simon. Smart pac-learners. *Theor. Comput. Sci.*, 412 (19):1756–1766, 2011.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *Journal of Machine Learning Research*, 1:245–279, 2001.

M. Kääriäinen. Generalization error bounds using unlabeled data. *COLT*, 2005.

Y. Li, P. M. Long, and A. Srinivasan. The one-inclusion graph algorithm is near-optimal for the prediction model of learning. *IEEE Transactions on Information Theory*, 47(3): 1257–1261, 2001.

P.A.P. Moran. The random division of an interval. *Supplement to the Journal of the Royal Statistical Society*, 9(1):92–98, 1947.

L. Pitt and M. K. Warmuth. Prediction preserving reducibility. *Journal of Computer and System Sciences*, 41(3), 1990.

R. Urner, S. Shalev-Shwartz, and S. Ben-David. Access to unlabeled data can speed up prediction time. *ICML*, 2011.

V. N. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and remote control*, 24, 1963.

X. Zhu and A. B. Goldberg. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan-Claypool, 2009.

Appendix A. Algebraic completion of the proof of Theorem 2

Here we give an algebraic proof that $\int_{s\in R_i} error(s)ds \le g_i^2/4$, whose proof was sketched using plots in Section 4. This integral corresponds to the situation where the shifted sample causes θ to fall in the "gap" between x_i and x_{i+1}. We again assume that $\theta \le 1/2$ (the other case is symmetrical) and proceed using the same cases. As before, let r_i be the shift taking x_{i+1} to θ and g_i be the length of the x_i, x_{i+1} gap. Thus $x_{i+1} \oplus r_i = \theta$ and a shift of $r_i + g_i$ takes x_i to θ even though $r_i + g_i$ might be greater than one.

Case (A): $g_i \le \theta$. In this case, θ falls in the gap only for shifts s where $x_i \oplus s \le \theta < x_{i+1} \oplus s$. The maximum margin algorithm makes a mistake on a randomly drawn test point exactly when the test point is between the middle of the (shifted) gap and θ. Therefore,

$$\int_{r_i}^{r_i+g_i} error(s)ds = \int_{r_i}^{r_i+g_i} \left| \left(\frac{x_i + x_{i+1}}{2}\right) \oplus s - \theta \right| ds = 2\int_0^{g_i/2} z\, dz = \frac{g_i^2}{4}.$$

Case (B): $\theta \le g_i \le 1-\theta$. In this case, as before, the integral goes from r_i to $r_i + g_i$. However, the expected error is slightly more complicated: while $s \in [r_i, r_i + g_i - \theta]$, all of the examples are negative, and the expected error is $\left|\frac{x_{i+1}\oplus s}{2} - \theta\right|$, and when $s \in [r_i+g_i-\theta, r_i+g_i]$ the expected error is $\left|\frac{x_{i+1}+x_i}{2} \oplus s - \theta\right|$, see the Case B plots in Section 4. Thus:

$$\int_{r_i}^{r_i+g_i} error(s)ds = \int_{r_i}^{r_i+g_i-\theta} \left| \frac{x_{i+1}\oplus s}{2} - \theta \right| ds + \int_{r_i+g_i-\theta}^{r_i+g_i} \left| \frac{x_{i+1}\oplus s + x_i \oplus s}{2} - \theta \right| ds.$$

Using a change of variables ($t = s - r_i$), we get

$$\begin{aligned} \int_{r_i}^{r_i+g_i} error(s)ds &= \int_0^{g_i-\theta} \left| \frac{\theta + t}{2} - \theta \right| dt + \int_{g_i-\theta}^{g_i} \left| \frac{\theta + t + \theta - g_i + t}{2} - \theta \right| dt \\ &= \int_0^{g_i-\theta} \left| \frac{t-\theta}{2} \right| dt + \int_{g_i-\theta}^{g_i} \left| t - \frac{g_i}{2} \right| dt. \end{aligned} \tag{10}$$

Subcase (B1): $g_i \ge 2\theta$. Continuing from Equation (10),

$$\begin{aligned} \int_{r_i}^{r_i+g_i} error(s)ds &= \int_0^{\theta} \frac{\theta - t}{2}\, dt + \int_{\theta}^{g_i-\theta} \frac{t-\theta}{2}\, dt + \int_{g_i-\theta}^{g_i} \frac{g_i}{2} - t\, dt \\ &= \frac{\theta^2}{2} - \frac{\theta^2}{4} - \frac{\theta(g_i - 2\theta)}{2} + \frac{(g_i-\theta)^2}{4} - \frac{\theta^2}{4} + \frac{\theta g_i}{2} - \frac{g_i^2}{2} + \frac{(g_i-\theta)^2}{2} \\ &= \frac{g_i^2}{4} + \frac{7\theta^2}{4} - \frac{3g_i\theta}{2} < g_i^2/4. \end{aligned}$$

Subcase (B2): $\theta < g_i < 2\theta$. Again continuing from Equation (10),

$$\begin{aligned} \int_{r_i}^{r_i+g_i} error(s)ds &= \int_0^{g_i-\theta} \frac{\theta - t}{2}\, dt + \int_{g_i-\theta}^{g_i/2} \frac{g_i}{2} - t\, dt + \int_{g_i/2}^{g_i} t - \frac{g_i}{2}\, dt \\ &= \frac{\theta(g_i-\theta)}{2} - \frac{(g_i-\theta)^2}{4} + (\theta - \frac{g_i}{2})\frac{g_i}{2} - \frac{g_i^2}{8} + \frac{(g_i-\theta)^2}{2} - \frac{g_i^2}{4} + \frac{g_i^2}{2} - \frac{g_i^2}{8} \\ &= \frac{2\theta g_i - \theta^2}{4} < \frac{g_i^2}{4} \end{aligned}$$

since $2\theta g_i - \theta^2$, as a function of θ, is nondecreasing on the interval $(g_i/2, g_i)$, and therefore at most g_i^2 over that interval.

Case (C): $g_i \geq 1 - \theta$. When $s \in [r_i, r_i + g_i - \theta]$, all of the examples are negative (as in case (B)) and when $s \in (r_i + 1 - \theta, r_i + g_i)$ all the examples are positive. This partitions the shifts in $(r_i, r_i + g_i)$ into three parts (see the plots in Section 4). Initially θ falls in the gap between 0 and the shifted x_{i+1}. Then x_i shifts in and the θ is in the gap between the shifted x_i and x_{i+1}. Finally, x_{i+1} wraps around and θ is in the gap between the shifted x_i and 1.

Thus $\int_{r_i}^{r_i+g_i} error(s)\ ds$ equals

$$\int_{r_i}^{r_i+g_i-\theta} \left|\frac{x_{i+1} \oplus s}{2} - \theta\right| ds + \int_{r_i+g_i-\theta}^{r_i+1-\theta} \left|\frac{x_{i+1} \oplus s + x_i \oplus s}{2} - \theta\right| ds + \int_{r_i+1-\theta}^{r_i+g_i} \left|\frac{x_i \oplus s + 1}{2} - \theta\right| ds$$

Using the substitution $t = s - r_i$ and following case B this becomes

$$\int_{r_i}^{r_i+g_i} error(s)\ ds = \int_0^{g_i-\theta} \left|\frac{t-\theta}{2}\right| dt + \int_{g_i-\theta}^{1-\theta} \left|t - \frac{g_i}{2}\right|\ dt + \int_{1-\theta}^{g_i} \left|\frac{\theta - g_i + t + 1}{2} - \theta\right|\ dt \quad (11)$$

Subcase (C1): $g_i \geq 2\theta$, so $1 - g_i \leq \theta \leq g_i/2$. Continuing from (11), $\int_{r_i}^{r_i+g_i} error(s)\ ds$ equals

$$\begin{aligned}
&\int_0^{\theta} \frac{\theta - t}{2}\ dt + \int_{\theta}^{g_i-\theta} \frac{t-\theta}{2}\ dt + \int_{g_i-\theta}^{1-\theta} t - \frac{g_i}{2}\ dt + \int_{1-\theta}^{g_i} \frac{t+1-g_i-\theta}{2}\ dt \\
&\qquad = \frac{\theta^2}{4} + \frac{(g_i - 2\theta)^2}{4} + \frac{(1-g_i)(1-2\theta)}{2} + \frac{4g_i(1-\theta) - g_i^2 - 3(1-\theta)^2}{4} \\
&\qquad = \frac{g + \theta^2 + \theta}{2} - g\theta - 1/4.
\end{aligned}$$

Note that the second derivative w.r.t. θ is positive, so the r.h.s. is maximized when $\theta = 1 - g_i$ or $\theta = g_i/2$. In both cases, it is easily verified that the value is at most $g_i^2/4$.

Subcase (C2): $g_i \leq 2\theta$. Continuing from (11) and following the logic of case B2, $\int_{r_i}^{r_i+g_i} error(s) ds$ equals

$$\begin{aligned}
&\int_0^{g_i-\theta} \frac{\theta - t}{2}\ dt + \int_{g_i-\theta}^{g_i/2} \frac{g_i}{2} - t\ dt + \int_{g_i/2}^{1-\theta} t - \frac{g_i}{2}\ dt + \int_{1-\theta}^{g_i} \frac{t+1-g_i-\theta}{2}\ dt \\
&\qquad = \frac{(3\theta - g_i)(g_i - \theta)}{4} + \frac{(2\theta - g_i)^2}{8} + \frac{(2 - g_i - 2\theta)^2}{8} + \frac{(g_i + \theta - 1)(3 - g_i - 3\theta)}{4} \\
&\qquad = \frac{2g_i + 2\theta - 1 - g_i^2 - 2\theta^2}{4}.
\end{aligned}$$

This is increasing in θ, and thus maximized when $\theta = 1/2$ where it becomes $g_i/2 - g_i^2/4 - 1/8$. We want to show that this bound is at most $g_i^2/4$, i.e. that $f(g_i) = g_i/2 - g_i^2/2 - 1/8 \leq 0$. Combining the facts that $g_i \geq 1/2$ in Case C, $f(1/2) = 0$, and $f'(g_i) \leq 0$ when $g_i \geq 1/2$, gives the desired inequality.