LA-UR- 95 - 0038

CONF-950478--1

Los Alamos National Laboratory is operated by the University of California for the United States Department of Energy under contract W-7405-ENG-36

TITLE: OFF-TRAINING-SET ERROR FOR THE GIBBS AND THE BAYES OPTIMAL GENERALIZERS

AUTHOR(S): Tal Grossman, T-13
Emanuel Knill, CIC-3
David Wolpert, Santa Fe Institute

SUBMITTED TO: ACM Conference on Computational Learning Theory, Santa Cruz, CA, July 1995

Los Alamos
Los Alamos National Laboratory
Los Alamos, New Mexico 87545

FORM NO. 836 R4
ST. NO. 2629 5/81

Off-Training-Set Error for the Gibbs and the Bayes Optimal Generalizers

Tal Grossman
Theoretical Division and CNLS, MS B213
LANL, Los Alamos, NM 87545
email: tal@goshawk.lanl.gov

Emanuel Knill
CIC 3
LANL, Los Alamos, NM 87545
email: knill@lanl.gov

David Wolpert
The Santa Fe Institute
1660 Old Pecos Trail, Suite A
Santa Fe, NM 87505
email: dhw@santafe.edu

January 3, 1995

Abstract

In this paper we analyze the average off-training-set behavior of the Bayes-optimal and Gibbs learning algorithms. We do this by exploiting the concept of *refinement*, which concerns the relationship between probability distributions. For non-uniform sampling distributions the expected off training-set error for both learning algorithms can rise with training set size. However we show in this paper that for uniform sampling and either algorithm, the expected error is a non-increasing function of training set size. For uniform sampling distributions, we also characterize the priors for which the expected error of the Bayes-optimal algorithm stays constant. In addition we show that when the target function is fixed, expected off-training-set error can increase with training set size if and only if the expected error averaged over all targets decreases with training set size. Our results hold for arbitrary noise and arbitrary loss functions.

1 Introduction

This paper is concerned with the supervised learning problem: There is an unknown *target* relationship f between an input space and an output space. A *training set* of i-o pairs is generated by sampling the target relationship. The problem is to use the training set to guess the input-output relationship which best fits the target relationship according to some suitable cost function. Such a guessed relationship from questions to outputs is known as a *hypothesis* relationship. An algorithm which produces a hypothesis relationship from a training set is called a *generalizer* or a *learning algorithm.*

Conventionally, a learning algorithm's performance is measured using test sets produced by the same process that generated the training set. Such an "iid" error function allows overlap between training and test sets. If one instead concentrates on off training-set (OTS) error, a set of "no-free-lunch" (nfl) theorems apply: averaged over all targets or averaged over all priors, all (fixed) generalizers perform the same [4]. However these theorems do not address the issue of how well a generalizer performs if it is coupled to the prior.

This paper is an investigation of this issue. We analyze several different aspects of the off-training-set behavior of Bayes-optimal and Gibbs learning algorithms [1, 2, 3, 4] in the case where the algorithms' assumption for the prior is correct.

One of the differences between our work and recent work by Haussler et al [1] is that we concentrate on OTS error learning curves averaging over both the input and output components of the training set. In contrast, the work of Haussler et al. deals either with the case where the training set inputs are fixed (while the outputs can vary), or, when those inputs are free to vary, considers the iid error function. Another important difference is that Haussler et al. restrict their attention to the noise-free case where the output space is binary and the loss function is the zero-one loss. In contrast, (most of) our results hold for arbitrary noise, arbitrary output spaces, and arbitrary loss functions.

Section 2 presents the mathematical framework we will use. It then presents a summary of why OTS error is of interest and briefly reviews the nfl theorems. In Section 3 we define the concept of a *refinement* and present some lemmas concerning refinement. As it turns out, OTS error can increase with training set size, on average, even for the Bayes-optimal algorithm (see [4].) In Section 4 we use our lemmas concerning refinement to show that this can not happen when the input space sampling distribution is uniform. We go on in that section to show under what conditions the OTS error of the Bayes optimal algorithm is constant and relate this to specific properties of the prior. We then prove similar results for the Gibbs algorithm. Full versions of proof sketches are provided in [5].

2 Preliminaries

2.1 The mathematical formalism and notation

This paper uses the extended Bayesian formalism ([4]. n and r are the number of elements in X (the input space) and Y (the output space) respectively. Most of our results are not limited to finite input and output spaces. However those cases are the simplest to present.

Our primary random variables are target relationships f, hypothesis relationships h, training sets d, and *cost* or *error* values c. In addition, it will be useful to relate these variables to one another using three other random variables. *Testing* (involved in determining the value of c) is done at the X value given by the random variable q. Y values associated with the hypothesis are denoted by y_h, and Y values associated with the target are denoted by y_f.

In this paper the target and the hypothesis relationships are defined by X-conditioned distributions over Y values. Formally:

$$f(q, y_f) \equiv P(y_f \mid f, q), \quad \text{and} \quad h(q, y_h) \equiv P(y_h \mid h, q).$$

If for each q, $h(q, y_h)$ is a delta function over y_h (i.e., if it specifies a single-valued function from X to Y), then h is called *single-valued.*

The value d of the training set random variable is an ordered set of m input-output examples. Those examples are indicated by $\{d_X(i), d_Y(i)\}_{i=1..m}$. The set of all input values in d is d_X and similarly for d_Y. The number of distinct values in d_X is denoted by m'. In the case of iid generation of d, d is formed by sampling X according to a *sampling distribution* $\pi(x)$ and then sampling f at the those points in X. More formally, the likelihood is

$$P(d \mid f) = \prod_{i=1}^{m} \pi\left(d_X(i)\right) f\left(d_X(i), d_Y(i)\right). \tag{1}$$

As an example, in the noise-free iid case, f is single-valued and $P(d \mid f)$ is given by $\delta(d \subseteq f) \prod_i \pi(d_X(i))$, where $\delta(d \subseteq f) = 1$ if d agrees with f, and 0 otherwise. In this paper though we are *not* assuming a noise free situation - our results hold for arbitrary f. Note that this means that two elements in the training set can have the same X values but different Y values. The *posterior* is the Bayesian inverse of the likelihood, $P(f \mid d)$.

Any learning algorithm is specified by $P(h \mid d)$. It is *deterministic* if the same d always gives the same h (i.e., if for fixed d $P(h \mid d)$ is a delta function about one particular h). In supervised learning $P(h \mid f, d) = P(h \mid d)$ (i.e., the learning algorithm only sees d in making its guess, not f). We will say that the full P "has" or "specifies" the generalizer given by $P(h \mid d)$.

In the case of iid error, $P(q \mid d) = \pi(q)$. In the case of OTS error,

$$P(q \mid d) = \frac{\pi(q)\delta(q \notin d_X)}{\sum_q \pi(q)\delta(q \notin d_X)}$$

where $\delta(z) = 1$ if z is true, 0 otherwise.

The cost c is defined by $c = L(y_h, y_f)$. For example, zero-one loss has $L(a, b) = 1 - \delta_{a,b}$, and quadratic loss has $L(a, b) = (a - b)^2$. It is easy to show that the expected cost given f, h and d is $E(c \mid f, h, d) = \sum_{y_h, y_f, q} L(y_h, y_f) h(q, y_h) f(q, y_f) P(q \mid d_X)$. Unless explicitly stated otherwise, from now on, whenever referring to cost, we are implicitly referring to OTS cost.

2.2 OTS Error and the nfl theorems

Since iid error allows the test set to overlap with the training set, it gives a generalizer credit for simply memorizing the training set. If one instead wishes to measure the *generalization* ability of the generalizer, it is appropriate to use OTS error. In addition, in a number of real-world scenarios, we are explicitly interested only in OTS error. For example, this is the case in much of protein structure prediction for drug design [4].

Unfortunately, for OTS error we have the "no-free-lunch theorems" [4] limiting the assumption-free utility of any learning algorithm. In particular, consider the uniform average, over all distributions $P(f)$, of the expected cost given m, $E(c \mid m) = \sum_{f,h,d} E(c \mid f, h, d)\ P(f, h, d \mid m)$. For the zero-one loss function, this is the same for all learning algorithms. So loosely speaking, there are "just as many" priors in which any algorithm A has superior behavior to that of any other algorithm B as vice-versa, for OTS error. Similar results hold for other loss functions, and for other conditioning events. In this paper we investigate the case where the uniform average over priors is relaxed, so that the learning algorithm and $P(f)$ are coupled.

2.3 The Gibbs and Bayes-optimal generalizers

The Bayes-optimal generalizer is the one that minimizes expected cost given d, $E(c \mid d)$. In other words it produces the best (as far as expectation value is concerned) guess one could, given the training set at hand. It is deterministic, with its hypothesis $h^\star$ given by

$$h^\star(x) = \text{argmin}_y W(x, y) \tag{2}$$

where $W(x, y) \equiv \int df L(f(x), y) P(f \mid d)$ (See [1, 2, 4]). (In cases of multiple minima of $W(x, y)$, any tie-breaking scheme will do.)

For the zero-one $L(.,.)$ and single-valued f (i.e., $P(f)$ that equal 0 for non-single-valued f), $h^\star(x) = \text{argmax}_y \sum_f \delta(f(x), y) P(f \mid d)$. In particular, if $P(f)$ is uniform across all single-valued f in some "target class" U and zero otherwise, and if $Y = \{0, 1\}$, then for any x, $h^\star(x) = 1$ if the number of f in U that are consistent with d and go through the point $\{x, 1\}$ exceeds the number of f in U that are consistent with d and go through $\{x, 0\}$.

A Gibbs generalizer is one that obeys

$$P(h \mid d) \propto P(d \mid f)|_{f=h} G(h) \tag{3}$$

where $G(h)$ is a distribution and the proportionality constant (sometimes called a "partition function") is set by normalization and only depends on d [1, 2, 3, 4]. A "correct" Gibbs generalizers is a Gibbs generalizer for which $G(h) = P(f)|_{f=h}$. (In this paper we restrict attention to correct Gibbs generalizers.)

Distributions in which $P(h \mid d)$ is the Bayes-optimal generalizer will be subscripted with "BO". For Gibbs generalizers we will subscript with "G". Note that in the absence of noise, both generalizers produce hypotheses that agree exactly with d.

It turns out that for some non-uniform sampling distributions $E_{BO}(c \mid m')$ is an increasing function of m' [4, 5] for all m'. The same is true for Gibbs generalizers. Note that in general, if $E_{BO}(c \mid m' = 0) = E_G(c \mid m' = 0)$, then since $E_G(c \mid m' = k) \geq E_{BO}(c \mid m' = k)$, if $E_{BO}(c \mid m')$ increases when m' goes from 0 to k, so must $E_G(c \mid m')$.

3 Distribution refinements

For our purposes, the crucial characteristic of the relationship between a distribution conditioned on training set size m and one conditioned on training set size $m + 1$ is whether the latter is a *refinement* of the former.

Definition The distribution P over the variables q, d, z is a refinement of the distribution $\hat{P}$ over the same variables if there exists a function T such that for all values of q, d, and z,

$$\hat{P}(q, d, z) = \sum_{(q', d'): T(q', d') = (q, d)} P(q', d', z). \tag{4}$$

Changing the choice of the variable z changes the meaning of refinement. Note that for any random variable z, if one distribution is a refinement of the other with variable list (q, d, z), then by marginalization the same is true with variable list (q, d).

An example of refinement is the case where $P(q \mid d)$ is iid, so that probabilities involving any training set d of size m can be found by summing probabilities involving those training sets of size $m + 1$ whose first m elements are d:

Lemma 1. For iid $P(q \mid d)$ and $z = y_f$, the distribution conditioned on training set size $m + 1$ is a refinement of the distribution conditioned on training set size m.

Proof sketch: Expanding in terms of f and plugging in our likelihood (see Eq. 1)

$$P(q, d, y_f \mid m) = \left[\pi(q) \prod_{i=1,m} \pi(d_X(i)) \right] \int df f(q, y_f) P(f) \prod_{i=1,m} f(d_X(i), d_Y(i)) \ . \tag{5}$$

To establish refinement for this scenario, we must show that

$$\sum_{d(m)} P(q, d, y_f \mid m) = P(q, d - d(m), y_f \mid m - 1) \ , \tag{6}$$

where "$d(m)$" means the m'th pair in d and "$d - d(m)$" is the first $m - 1$ pairs in d. (Here $T(q', d') \equiv (q', (d' - d'(m))$.) By normalization, $\sum_y f(x, y) = 1$, and $\sum_x \pi(x) = 1$. So breaking the $\sum_{d(m)}$ into first a sum over $d_Y(m)$ and then one over $d_X(m)$ establishes refinement. **QED**

For OTS error, although $P(q \mid d, f) = P(q \mid d)$ as in the derivation of lemma 1, we no longer have $P(q \mid d) = \pi(q)$. Plugging in the OTS $P(q \mid d)$, we obtain $P(d, q, y_f \mid m) =$

$$\frac{\pi(q)\delta(q \notin d_X)}{\sum_q \pi(q)\delta(q \notin d_X)} \times \left[\prod_{i=1,m} \pi(d_X(i))\right] \int df f(q, y_f) P(f) \prod_{i=1,m} f(d_X(i), d_Y(i)) \ . \tag{7}$$

As before, to establish refinement we must prove (6). Whether it holds depends on $\pi(x)$. In particular, there are non-uniform $\pi(x)$ for which it doesn't hold (since there are such $\pi(x)$ for which expected error increases with m and, as shown below, refinement implies non-increasing expected error). However we have the following special case:

Lemma 2. For OTS $P(q \mid d)$ and $z = y_f$, the distribution conditioned on training set size $m + 1$ is a refinement of the distribution conditioned on training set size m, if $\pi(x)$ is uniform.

Proof sketch: Dropping the subscript f on y, from (7) we get

$$P(d, q, y \mid m) = \frac{n^{-m}}{(n - m')} \prod_{i=1,m} \delta(q \neq d_X(i)) \times \int df f(q, y) P(f) \prod_{i=1,m} f(d_X(i), d_Y(i)) \ .$$

A similar expression holds for $P(d - d(m), q, y \mid m - 1)$. We want to sum $P(d, q, y \mid m)$ over all $d(m)$ to get that expression for $P(d - d(m), q, y \mid m - 1)$. First summing $P(d, q, y \mid m)$ over all $d_Y(m)$, we get

$$\frac{n^{-m}}{(n - m')} \prod_{i=1,m} \delta(q \neq d_X(i)) \times \int df f(q, y) P(f) \prod_{i=1,m-1} f(d_X(i), d_Y(i)) \ .$$

where "$m'(-1)$" means the number of distinct elements in the first $(m - 1)$ elements of d.

Now expand

$$\frac{n^{-m}}{(n - m')} \prod_{i=1,m} \delta(q \neq d_X(i)) = \frac{n^{-m}}{(n - m')} \prod_{i=1,m-1} \delta(q \neq d_X(i)) \{\delta(q \neq d_X(m))\} \ .$$

The only thing affected by the value of $d_X(m)$ is the $\{.\}$ term and m'. Moreover, it is not hard to show that $\sum_{d_X(m)} \{.\}/(n - m') = n/(n - m'(-1))$. This completes the proof sketch. **QED**

4 Refinement and Generalization

It is not hard to show that

$$E(c \mid m) = \sum_{y_f, y_h, d, q} L(y_f, y_h) P(y_h \mid d, q) P(y_f, d, q \mid m). \tag{8}$$

Now $E(c \mid m)$ is minimized by the Bayes-optimal generalizer [4]. Given Eq. 8, this implies that for any distribution P, if P_{BO} has a Bayes-optimal generalizer, then

$$\sum_{d, y_h, y_f, q} L(y_h, y_f) P_{BO}(y_f, q, d) P_{BO}(y_h \mid d, q) \leq \sum_{d, y_h, y_f, q} L(y_h, y_f) P_{BO}(y_f, q, d) P(y_h \mid d, q) \tag{9}$$

(the conditioning on m being implicit). This is the underlying reason why, in many scenarios, expected cost shrinks as training set size increases for the Bayes-optimal generalizer. Formally:

Theorem 1. If P_2 is a refinement of P_1 with $z = y_f$ and if both P_1 and P_2 have Bayes-optimal generalizers, then $E_2(C) \leq E_1(C)$, where the subscript indicates the distribution.

Proof sketch: We can replace $P_1(y_f, d, q)$ with $\sum_{(q', d'): T(q', d') = (q, d)} P_2(q', d', y_f)$, because P_2 is a refinement of P_1. Doing this in the expansion for $E_1(c)$ gives

$$E_1(c) = \sum_{y_h, y_f, q, d, q', d': T(q', d') = (q, d)} P_2(q', d', y_f) P_1(y_h \mid d, q) L(y_h, y_f) \ .$$

Substituting in $T(d')$ for d and $T(q')$ for q and then relabeling, we get

$$E_1(c) = \sum_{y_h, y_f, q', d', q, d: T(q, d) = (q', d')} P_2(q, d, y_f) P_1(y_h \mid T(d), T(q)) L(y_h, y_f) \ .$$

Now for any q and d there is a q' and a d' such that $T(q, d) = (q', d')$. Therefore for any (!) q and d, if we check all q' and d' we will find one such pair for which $T(q, d) = (q', d')$. Therefore our sum must extend over all q and d. Accordingly, we can rewrite our sum as

$$E_1(c) = \sum_{y_h, y_f, q, d} P_2(q, d, y_f) P_1(y_h \mid T(d), T(q)) L(y_h, y_f) \ .$$

By Bayes-optimality, this is greater or equal to

$$E_2(c) = \sum_{y_h, y_f, q, d} P_2(q, d, y_f) P_2(y_h \mid d, q) L(y_h, y_f). \qquad \textbf{QED}$$

By this theorem and the refinement lemmas, if either a) we have iid error; or b) we have OTS error and a uniform sampling distribution; then $E(c \mid m) \geq E(c \mid m + 1)$ for all m for the Bayes optimal generalizer. Note that this result holds for any $P(d_y \mid d_x, f)$ (i.e., any noise process) and any loss function L.

4.1 Context of these results

It is important to note that for non-uniform $\pi(x)$, $E(c \mid m)$ may increase with m (an example is given in [4]). In fact, even more "intuitively obvious" results than that of theorem (1) can fail to hold for OTS error when $\pi(x)$ is non-uniform. For example, it is proven in [5] that

Lemma 3. For uniform $\pi(x)$ and $P(h \mid d) = \delta(h, g(x))$ for some function $g(x)$, $E(c \mid m')$ is independent of m'.

It may seem obvious that the expected cost of the generalizer that always guesses the same function, no matter what the data, doesn't vary with the size of that data. However this is *not* true in general for non-uniform $\pi(x)$ and OTS error. An example is the scenario presented in [4] for which the OTS error of the Bayes optimal generalizer increases with m'.

The fact that π's being uniform is important both for theorem (1) and lemma (3) is no coincidence; the two results are closely related, as the following lemma (proven in [5]) shows:

Lemma 4. For uniform $\pi(x)$ and $0 < k < n$, $E_{BO}(c \mid m' = k) = E_{BO}(c \mid m' = 0)$ iff there exists a data-independent $g(x)$ such that $E_{BO}(c \mid m' = k) = E_g(c \mid m' = k)$.

Any generalizer that minimizes $E(c \mid m)$ necessarily makes the same guesses as the Bayes-optimal generalizer for all d such that $P(d) \neq 0$ [4]. (Recall that if more than one h minimize $E(c \mid d)$ for some d, then saying a generalizer is "Bayes-optimal" simply means that it guesses one of those optimal h in response to d.) Accordingly, lemma (4) tells us that $E_{BO}(c \mid m' = k) = E_{BO}(c \mid m' = 0)$ if and only if one Bayes-optimal generalizer guesses some function $g(x)$ in response to all allowed training sets.

4.2 Constant OTS error and Ideals

An interesting problem is to determine the conditions under which the OTS error is constant for all $m' < n$ (so that OTS error is defined). We analyze this problem for the simple case where Y is binary, there is no noise, and $\pi(x)$ is uniform. In this case, all targets are vectors living at the vertices of the hypercube $2^{[n]}$. In addition, the only d such that $P(d) = 0$ are those that lie on an f for which $P(f) = 0$. For $u, v \in 2^{[n]}$, let $u \oplus v$ denotes the pointwise exclusive OR between u and v. Let $u \geq v$ mean that $u_i \geq v_i$ for every i.

Definition Let P be a probability distribution on $2^{[n]}$. P is a *relative ideal* iff there exists a $g \in 2^{[n]}$ (the *center* of P) such that for all $u, v \in 2^{[n]}$ with $u \oplus g \geq v \oplus g$, $P(u) \leq P(v)$.

A simple example of a relative ideal is a P that is constant for all vectors within some Hamming distance R of g, and zero otherwise.

Theorem 2 $E_{BO}(c|m')$ is constant for all m' up to $m' = n - 1$ iff $P(f)$ is a relative ideal.

Proof sketch: Suppose that $E_{BO}(c|m')$ is constant for all m' up to $m' = n - 1$. Then by the discussion following lemma (4), one Bayes optimal generalizer (there may be more than one) guesses the same function $g \in 2^{[n]}$ in response to any allowed training set having $m' < n$. Suppose that t is the function defined by a training set d having $m' = n - 1$. (The domain of t is d_X.) Let i be the only input not seen. The only targets consistent with d are the two extensions of t defined by $t_i^{(0)} \equiv g_i$ and $t_i^{(1)} \equiv g_i \oplus 1$. Since g is a Bayes optimal hypothesis, $P(f)|_{f=t^{(1)}} \leq P(f)|_{t^{(0)}}$. This inequality holds no matter what t is (so long as it lies on an f with non-zero prior probability), and no matter what input remains to be seen.

Now suppose that $u \oplus g \geq v \oplus g$, and that $P(f)|_{f=u} \neq 0$. (If $P(f)|_{f=u} = 0$, then it trivially follows that $P(f)|_{f=u} \leq P(f)|_{f=v}$.) Since $P(f)|_{f=u} \neq 0$, any training set lying on u is allowed. Furthermore, since $u \oplus g \geq v \oplus g$, there is a chain $v_0 = u, v_1, \ldots, v_k = v$ such that $v_i \oplus g \geq v_{i+1} \oplus g$ and v_i and v_{i+1} differ in only one position. This implies that $P(f)|_{f=v_i} \leq P(f)|_{f=v_{i+1}}$ for each i (use the argument of the preceding paragraph with t defined to be the $n - 1$ points on which v_i and v_{i+1} agree). Hence, $P(f)|_{f=u} \leq P(f)|_{f=v}$.

To prove the other direction, suppose that P is a relative ideal with center g and that we are given a d with $m' = n - 1$. As in the previous paragraph, let t be the function defined by d, let i be the sole element of X not in d_X, and define the two extensions of t by $t_i^{(0)} \equiv g_i$ and $t_i^{(1)} \equiv g_i \oplus 1$. The inequality of the definition of relative ideals implies that g is a Bayes optimal guess for input i (choose $v = t^{[0]}$ and $u = t^{[1]}$). Since this is true for any d, it is also true when we average over d's. The inequality also implies that g is the Bayes optimal guess when no training examples have been seen (choose $v = g$). So the Bayes-optimal generalizer is identical to the generalizer of lemma (3) for these two m' values, for points outside of the training set. Accordingly, by lemma (3), $E_{BO}(c|m')$ is the same for those two m' values. Therefore by theorem (1), it is independent of m'. **QED**

Next consider $E_{BO}(c \mid f^*, m')$, where f^* is some target with non-zero prior probability. Even for a uniform $\pi(x)$ this expected cost can increase with increasing m' [4]. However the following corollary of theorem (2) allows us to rule out such behavior in many cases.

Corollary 1. For a uniform $\pi(x)$, for any target f^* whose prior probability does not equal zero, $E_{BO}(c \mid f^*, m')$, can increase with m' only if $P(f)$ is not a relative ideal.

Proof sketch: By theorem (2), if $P(f)$ is a relative ideal, $E_{BO}(c \mid m')$ is a constant function of m'. By lemma (4), that would imply that a Bayes-optimal generalizer (there may be more than one) always guesses the same function $g(x)$ in response to any allowed training set. This in turn would mean that the value of $E_{BO}(c \mid f^*, m')$ for the actual (relative ideal) prior at hand is equal to the value of $E_g(C \mid m')$ under the prior $P(f) = \delta(f - f^*)$. By lemma (3) however this latter expression can not increase with m'. **QED**

Now $E_{BO}(c \mid m') = \sum_f E_{BO}(c \mid f, m')P(f)$. So by Cor. (1) and Thm. (2), if $E_{BO}(c \mid m')$

does not vary with m', it has the same value as $E_{BO}(c \mid f, m')$ for any of its "constituent" f's. In this sense it has no variability over f. The inverse is as easily established: $E_{BO}(c \mid m')$ is a decreasing function of m' iff there is variability over f in the value of $E_{BO}(c \mid f, m')$.

4.3 Refinement and Gibbs generalizers

We can also use refinement to infer off-training set behavior for correct Gibbs generalizers.

Theorem 3. Let P_1 and P_2 have Gibbs generalizers where P_2 is a refinement of P_1. If $L(y, y')$ is a non-positive definite matrix over the subspace perpendicular to $\vec{1}$, then $E_2(C) \leq E_1(C)$.

Proof: Define $v_i(y_f, d, q) = P_i(y_f \mid d, q)$, where i can be 1 or 2. Define t as the pair (d, q). Now

$$E_i(C) = \sum_{y,y',t} L(y, y') v_i(y, t) v_i(y', t) P_i(t),$$

for a Gibbs generalizer. (Recall that $P(y_f \mid y_h, d, q) = P(y_f \mid d, q)$; see (8).) By refinement,

$$v_1(y, t) = \frac{1}{P_1(t)} \sum_{t' : T(t') = t} v_2(y, t') P_2(t') \ .$$

Replace $P_2(t')$ in the summand with $P_2(t', T(t') = t)$. By refinement, $P_1(t) = \sum_{t' : T(t') = t} P_2(t') \equiv P_2(t' : T(t') = t)$. Therefore, with $w(t', t) \equiv P_2(t' \mid T(t') = t)$,

$$E_1(C) = \sum_{y,y',t,t",t" : T(t') = t, T(t") = t} L(y, y') P_1(t) v_2(y, t') v_2(y', t") w(t', t) w(t", t) \ .$$

We want to express $E_2(C)$ as a similar sum. To that end, write

$$E_2(C) = \sum_{y,y',t} L(y, y') v_2(y, t) v_2(y', t) P_2(t) \ .$$

Now $P_2(t) = \sum_{t' : T(t) = t'} w(t, t') P_2(t : T(t) = t') = \sum_{t' : T(t) = t'} w(t, t') P_1(t')$. Relabeling,

$$\begin{aligned} E_2(C) &= \sum_{y,y',t,t' : T(t') = t} L(y, y') v_2(y, t') v_2(y', t') P_1(t) w(t', t) \\ &= \sum_{y,y',t,t',t" : T(t') = t, T(t") = t} L(y, y') P_1(t) v_2(y, t') v_2(y', t') w(t', t) w(t", t) \end{aligned} \tag{10}$$

The only difference between the expression for $E_1(C)$ and $E_2(C)$ is whether the summand contains $v_2(y', t")$ or $v_2(y', t')$. It is this difference that establishes the theorem. Write

$$E_2(C) - E_1(C) = \sum_{y,y',t,t',t" : T(t') = t, T(t") = t} L(y, y') P_1(t) w(t', t) w(t", t) v_2(y, t') \left[v_2(y', t') - v_2(y', t") \right] \ .$$

Now rewrite this expression by interchanging t' and $t"$. Add our two expressions for $E_1(C) - E_2(C)$ and divide by 2. The result is $E_2(C) - E_1(C) =$

$$\sum_{y,y',t,t',t":T(t')=t,T(t")=t} P_1(t)w(t',t)w(t",t)\frac{L(y,y')}{2}\left[v_2(y,t')-v_2(y,t")\right]\left[v_2(y',t')-v_2(y',t")\right] \quad .$$

Since it is the difference of two probability distributions, each [..] term, considered as a vector indexed by y (or y' as the case may be), is perpendicular to the unit vector. Given that L is non-positive definite by hypothesis, we see that $E_2(C) \leq E_1(C)$. **QED**

Both the zero-one and the quadratic L's have the required property. Intuitively, those loss functions for which the Gibbs generalizer's learning curve increases are those for which the loss shrinks as h and f get *further* apart. Interestingly, the learning curve is non-increasing for the Bayes-optimal generalizer even for such a "backwards" loss function.

5 Discussion

In this paper we introduce the refinement concept. We use it to prove that for a uniform sampling distribution $\pi(x)$, for any loss function and any noise process, the learning curve for OTS error for a Bayes optimal generalizer is non-increasing. (It can increase for non-uniform $\pi(x)$.) We also characterize those priors for which the learning curve is constant. We also use refinement to prove that for a uniform $\pi(x)$ the Gibbs generalizer has a non-increasing learning curve, provided certain (common) conditions on the loss function are met.

There are many questions that this analysis raises. Examples are: i) What is the behavior of $E_g(c \mid m) - E_{BO}(c \mid m)$? ii) What are the widths (as one varies f, varies d, etc.) of the distributions whose means are examined above? iii) For non-uniform $\pi(x)$, is it possible for $E(c \mid m)$ to increase for one region of values of m and decrease for another, for a fixed $P(f)$?

References

[1] Haussler, D., Kearns, M., Schapire, R., Machine Learning, **14**, pp. 83-115, 1994.

[2] Opper, M., and Haussler D., in *Proceedings of the 4th annual workshop on Computational Learning Theory*, 75–87. Morgan Kaufmann, 1991.

[3] Tishby, N., Levin E. and Solla S., in *International Joint Conference on Neural Networks*, **II**, 403–409. IEEE, 1989.

[4] Wolpert, D. in *The Mathematics of Generalization*, D. Wolpert Ed., 117–214, Addison-Wesley, 1994.

[5] Wolpert, Knill, Grossman, (1994). An Investigation of Off-Training-Set Behavior for the Gibbs and the Bayes Optimal Learning Algorithms, in preparation.