Journal of Machine Learning Research 4 (2003) 895-911 Submitted 3/03; Published 10/03

Concentration Inequalities for the Missing Mass and for Histogram Rule Error

David McAllester MCALLESTER@TTI-C.ORG
Toyota Technological Institute at Chicago
1427 East 60th Street
Chicago Il, 60637

Luis Ortiz LEORTIZ@LINC.CIS.UPENN.EDU
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

Editors: Ralf Herbrich and Thore Graepel

Abstract

This paper gives distribution-free concentration inequalities for the missing mass and the error rate of histogram rules. Negative association methods can be used to reduce these concentration problems to concentration questions about independent sums. Although the sums are independent, they are highly heterogeneous. Such highly heterogeneous independent sums cannot be analyzed using standard concentration inequalities such as Hoeffding's inequality, the Angluin-Valiant bound, Bernstein's inequality, Bennett's inequality, or McDiarmid's theorem. The concentration inequality for histogram rule error is motivated by the desire to construct a new class of bounds on the generalization error of decision trees.

1. Introduction

The Good-Turing missing mass estimator was developed in the 1940s to estimate the probability that the next item drawn from a fixed distribution will be an item not seen before. Since the publication of the Good-Turing missing mass estimator in 1953 (Good, 1953), this estimator has been used extensively in language modeling applications (Chen and Goodman, 1998, Church and Gale, 1991, Katz, 1987). Recently a large deviation accuracy guarantee was proved for the missing mass estimator (McAllester and Schapire, 2000, Kutin, 2002). The main technical result is that the missing mass itself concentrates—McAllester and Schapire (2000) prove that the probability that missing mass deviates from its expectation by more than ε is at most $e^{-m\varepsilon^2/3}$ independent of the underlying distribution. Here we give a simpler proof of the stronger result that the deviation probability is bounded by $e^{-m\varepsilon^2}$.

A histogram rule is defined by two things—a given clustering of objects into classes and a given training sample. In a classification setting the histogram rule defined by a given clustering and sample assigns to each cluster the label that occurred most frequently for that cluster in the sample. In a decision-theoretic setting, such as that studied by Ortiz and Kaelbling (2000), the rule associates each cluster with the action choice of highest performance on the training data for that cluster. We show that the performance of a histogram rule (for a fixed clustering) concentrates near its expectation—the probability that the performance deviates from its expectation by more than ε is bounded by $e^{-m\varepsilon^2/9}$ independent of the clustering or the underlying data distribution.

The concentration inequality for histogram rule error can be motivated by the desire to construct a new class of bounds on the generalization error of decision trees. A decision tree can be viewed as defining both a data clustering and a label for each cluster. It is possible to give highly compact specifications of tree structure (data clusterings). The concentration inequality for histogram rule error can be used to give a generalization bound on decision trees in terms of the number of bits needed to specify the tree structure independent of the number of leaves. For example, a compactly specified tree structure might have an infinite number of leaves. This potential application of the concentration inequality for histogram rule error is discussed in detail in the future work section (Section 10).

In proving the concentration inequalities for the missing mass and histogram rule error this paper makes use of a lemma relating "Chernoff Entropy" to the variance of the Gibbs distribution for an arbitrary real-valued random variable (an arbitrary configuration function). This general Gibbs-variance lemma is implicit in Section 6 of Chernoff's classic paper (Chernoff, 1952) but its utility seems to have gone unnoticed. Although all of the results in this paper can be proved without the Gibbs-variance lemma, this lemma is convenient in many cases. Section 3 shows how Hoeffding's inequality, the Angluin-Valiant bounds, a form of Bernstein's inequality, and a form of Bennett's inequality can all be viewed as direct corollaries of the general Gibbs-variance lemma.

Negative association results can be used to reduce the concentration problems for the missing mass and histogram rule error to concentration questions about independent sums. Section 6 shows that Hoeffding's inequality, the Angluin-Valiant bound, Bernstein's inequality and Bennett's inequality are all inadequate for the extremely heterogeneous independent sum problem underlying the missing mass. Section 7 then derives a concentration inequality for the missing mass directly from the Gibbs-variance lemma. Section 8 derives an incomparable concentration inequality for the missing mass from a lemma of Kearns and Saul (1998). Section 9 then derives a concentration inequality for histogram rule error from the Kearns-Saul lemma and convenient but inessential use of the Gibbs-variance lemma.

2. The Exponential Moment Method

The main technical challenge in proving the results of this paper is the analysis of extremely heterogeneous independent sums. It is possible to show that standard concentration inequalities for independent sums (Hoeffding's inequality, the Angluin-Valiant bound, Bernstein's inequality, and Bennett's inequality) are insufficient for extreme heterogeneity. Hence we must go back to the underlying exponential moment method used to prove the standard inequalities.

Let X be any real-valued random variable with finite mean. Let $DP(X,x)$ be $P(X \geq x)$ if $x \geq \mathrm{E}[X]$ and $P(X \leq x)$ if $x < \mathrm{E}[X]$. The following lemma is the central topic of Chernoff's classic paper (Chernoff, 1952).

Lemma 1 (Chernoff) *For any real-valued variable X with finite mean $\mathrm{E}[X]$ we have the following for any x where the "entropy" $S(X,x)$ is defined as below.*

$$DP(X,\, x) \quad \leq \quad e^{-S(X,x)} \tag{1}$$

$$S(X,x) = \sup_{\beta} x\beta - \ln Z(X,\, \beta) \tag{2}$$

$$Z(X,\, \beta) = \mathrm{E}\left[e^{\beta X}\right] \tag{3}$$

Lemma 1 follows, essentially, from the observation that for $\beta \geq 0$ we have the following.

$$P(X \geq x) \leq \mathrm{E}\left[e^{\beta(X-x)}\right] = e^{-\beta x}\mathrm{E}\left[e^{\beta X}\right] = e^{-(x\beta - \ln Z(X,\, \beta))} \tag{4}$$

Lemma 1 is called the exponential moment method because of the first inequality in (4).

In this paper we use some further general observations about the exponential moment method. For any real-valued random variable X there exists a unique largest open interval $(\beta_{\min},\, \beta_{\max})$ (possibly with infinite endpoints) such that for $\beta \in (\beta_{\min}, \beta_{\max})$ we have that $Z(X,\beta)$ is finite. For a discrete distribution, and for $\beta \in (\beta_{\min},\, \beta_{\max})$, the Gibbs distribution P_β can be defined as follows.

$$P_\beta(X = x) = \frac{1}{Z(X,\, \beta)} P(X = x)e^{\beta x}$$

For $\beta \in (\beta_{\min},\, \beta_{\max})$ we define the expectation of $f(X)$ at inverse temperature β as follows.

$$\mathrm{E}_\beta\left[f(X)\right] = \frac{1}{Z(X,\, \beta)}\ \mathrm{E}\left[f(X)e^{\beta X}\right]$$

The distribution P_β and the expectation operator $\mathrm{E}_\beta[\cdot]$ are both easily generalized to continuous distributions.

For $\beta \in (\beta_{\min},\, \beta_{\max})$ let $\sigma^2(X,\, \beta)$ be $\mathrm{E}_\beta\left[(X - \mathrm{E}_\beta[X])^2\right]$. The quantity $\sigma^2(X,\, \beta)$ is the Gibbs-variance at inverse temperature β. For $\beta \in (\beta_{\min}, \beta_{\max})$ we let $KL(P_\beta||P)$ denote the KL-divergence from P_β to P which can be written as follows where the first line is the general definition of KL-divergence.

$$\begin{aligned}
KL(P_\beta||P) &= \mathrm{E}_\beta\left[\ln \frac{dP_\beta(X)}{dP(X)}\right] \\
&= \mathrm{E}_\beta\left[\ln \frac{e^{\beta X}}{Z}\right] \\
&= \mathrm{E}_\beta[\beta X] - \mathrm{E}_\beta[\ln Z] \\
&= \mathrm{E}_\beta[X]\,\beta - \ln Z
\end{aligned} \tag{5}$$

Let $x_{\min}$ be the greatest lower bound of the set of all values of the form $\mathrm{E}_\beta[X]$ for $\beta \in (\beta_{\min},\, \beta_{\max})$ and let $x_{\max}$ be the least upper bound of this set. If the open interval $(x_{\min},\, x_{\max})$ is not empty then $\mathrm{E}_\beta[X]$ is a monotonically increasing function of $\beta \in (\beta_{\min},\, \beta_{\max})$. For $x \in (x_{\min},\, x_{\max})$ define $\beta(x)$ to be the unique value β satisfying $\mathrm{E}_\beta[X] = x$. Finally we define a definite double integral notation.

Definition 2 *For any continuous function f we define the definite double integral $\iint_a^x f(s)\, d^2s$ to be $F(x)$ where F is the unique function satisfying $F(a) = 0$, $F'(a) = 0$, and $F''(x) = f(x)$ where $F'(x)$ and $F''(x)$ are the first and second derivatives of F respectively.*

To calculate $\iint_a^b f(z)d^2z$ we can use $\iint_a^b f(z)d^2z = F(b) - F(a) - F'(a)(b-a)$ where $F(x)$ is the indefinite integral $\iint f(z)d^2z$, i.e., any function F with $F''(x) = f(x)$. For ε small we have the following.

$$\iint_a^{a+\varepsilon} f(z)d^2z \approx \frac{1}{2}f(a)\varepsilon^2$$

For any doubly differentiable function f we can write f as follows.

$$f(x) = f(a) + f'(a)(x-a) + \iint_a^x f''(z)d^2z \tag{6}$$

We now have the following general theorem.

Lemma 3 (Gibbs-variance lemma) *For any real-valued variable X, any $x \in (x_{\min}, x_{\max})$, and $\beta \in (\beta_{\min},\ \beta_{\max})$ we have the following.*

$$S(X,\ x) = x\beta(x) - \ln Z(X,\ \beta(x)) \tag{7}$$

$$= KL(P_{\beta(x)}||P) \tag{8}$$

$$= \iint_{\mathrm{E}[X]}^x \frac{d^2z}{\sigma^2(X,\ \beta(z))} \tag{9}$$

$$\ln Z(X,\ \beta) = \mathrm{E}_0\,[X]\,\beta + \iint_0^\beta \sigma^2(X,\ \gamma)d^2\gamma \tag{10}$$

Before proving this lemma we first discuss its significance. It is important to note that the theorem makes no assumptions about X—there are no boundedness or independence assumptions of any form. Furthermore, Equation (9) implies that for ε small we have the following.

$$S(X, \mathrm{E}_0\,[X] + \varepsilon) \approx \frac{\varepsilon^2}{2\sigma^2(X,\ 0)}$$

In any bound of the form $DP(X,\ \mathrm{E}_0\,[X] + \varepsilon) \leq e^{-f(\varepsilon)}$ proved by the exponential moment method we must have the following for small ε.

$$f(\varepsilon) \leq S(X,\ \mathrm{E}\,[X] + \varepsilon) \approx \frac{\varepsilon^2}{2\sigma^2(X,\ 0)}$$

So $f(\varepsilon)$ must be quadratic (or higher order) in ε and the constant of the quadratic term can not exceed $1/2\sigma^2$.

An interesting aspect of Lemma 3 is the connection it establishes between the general exponential moment method and general concepts of statistical mechanics. Up to sign conventions (7) is the standard statistical mechanics relation between the energy x, the entropy S, the inverse temperature β, and the Gibbs free energy $(\ln Z)/\beta$. Here we have that S and β have the opposite sign from standard thermodynamic conventions. The sign convention adopted here makes S positive and causes the sign of β to match to the sign of the deviation $\mathrm{E}_\beta\,[X] - \mathrm{E}_0\,[X]$.

3. Some Standard Concentration Inequalities

Although Lemma 3 makes no assumptions on X, it can be used in deriving more specialized bounds for, say, independent sums of bounded variables. Consider $X = \sum_i X_i$ where the X_i are independent and each X_i is bounded to an interval of width b_i. In this case we have $\sigma^2(X_i, \beta) \leq b_i^2/4$. It follows that $\sigma^2(X, \beta) \leq (1/4)\sum_i b_i^2$. Formula (9) then immediately implies the following.

$$S(X,\, \mathrm{E}_0[X] + \varepsilon) \geq \frac{2\varepsilon^2}{\sum_i b_i^2} \tag{11}$$

Combining this with (1) yields the following.

$$DP(X,\, \mathrm{E}_0[X] + \varepsilon) \leq e^{-2\varepsilon^2/\left(\sum_i b_i^2\right)} \tag{12}$$

Formula (12) is the familiar Hoeffding inequality (Hoeffding, 1963). We will generally write bounds in the form of (11) rather than the more familiar (12) because of the interaction of these bounds with other lemmas, e.g., Lemma 4 at the end of this section or the lemmas in Section 5.

Next consider $X = \sum_i X_i$ where the X_i are independent Bernoulli variables. For $\beta \leq 0$ we have the following.

$$\begin{aligned} \sigma^2(X,\, \beta) &= \sum_i P_\beta(X_i = 1)\left(1 - P_\beta(X_i = 1)\right) \leq \sum_i P_\beta(X_i = 1) \\ &\leq \sum_i P_0(X_i = 1) = \sum_i \mathrm{E}_0[X_i] = \mathrm{E}_0[X] \end{aligned}$$

Formula (9) then yields the following.

$$S(X,\, \mathrm{E}_0[X] - \varepsilon) \geq \frac{\varepsilon^2}{2\mathrm{E}_0[X]} \tag{13}$$

This is the lower deviation Angluin-Valiant bound (Angluin and Valiant, 1979, Hagerup and Rüb, 1989). A form of the upper-deviation Angluin-Valiant can be derived using the following observation.

$$\begin{aligned} \sigma^2(X,\, \beta) &= \sum_i P_\beta(X_i = 1)\left(1 - P_\beta(X_i = 1)\right) \leq \sum_i P_\beta(X_i = 1) \\ &\leq \mathrm{E}_\beta[X] \\ \sigma^2(X,\, \beta(x)) &\leq x \end{aligned} \tag{14}$$

Combining (9) with (14) yields the following.

$$S(X,\, \mathrm{E}_0[X] + \varepsilon) \geq \iint_{\mathrm{E}_0[X]}^{\mathrm{E}_0[X]+\varepsilon} \frac{d^2 z}{z} \tag{15}$$

$$= \mathrm{E}_0[X]\, H\left(\frac{\varepsilon}{\mathrm{E}_0[X]}\right) \tag{16}$$

where

$$H(t) = \iint_1^{1+t} \frac{1}{x}\, d^2 x = (1+t)\ln(1+t) - t$$

Formula (16) is Bennett's inequality for the Poisson limit where each p_i is very small so that $\sigma^2(X,\,0)$ is essentially equal to $\mathrm{E}_0[X]$. Note that in (16) we have that ε can be either positive or negative. It is interesting to note that in the Poisson limit formula (15) holds with equality and hence Bennett's inequality corresponds to the exact Chernoff entropy for the number of poison arrivals. This observation implies that (16) is the tightest possible concentration inequality provable by the exponential moment method for $DP(X,\,\mathrm{E}_0[X]+\varepsilon)$ for the case $X=\sum_i X_i$ with X_i independent and Bernoulli provided that we require the bound be a function of ε and $\mathrm{E}_0[X]$. For $t \geq 0$ it is possible to show that $H(t) \geq t^2/(2+(2/3)t)$ and (16) then yields the following.

$$S(X,\,\mathrm{E}_0[X]+\varepsilon) \geq \frac{\varepsilon^2}{2(\mathrm{E}_0[X]+\frac{1}{3}\varepsilon)} \tag{17}$$

Equation (17) is Bernstein's inequality in the case of the Poisson limit. For the case of $\varepsilon \leq \mathrm{E}_0[X]$ formula (17) implies the following.

$$S(X,\,\mathrm{E}_0[X]+\varepsilon) \geq \frac{\varepsilon^2}{\frac{8}{3}\mathrm{E}_0[X]} \tag{18}$$

Formula (18) is the upper deviation Angluin-Valiant bound although the bound is usually stated with the constant 3 replacing 8/3.

The following lemma implies that all of the inequalities discussed above generalize to independent sums of variables bounded to $[0,1]$.

Lemma 4 *Let $Y=\sum_i Y_i$ with Y_i independent and $Y_i \in [0,1]$. Let $X=\sum_i X_i$ with X_i independent and Bernoulli and with $\mathrm{E}[X_i]=\mathrm{E}[Y_i]$. For any such X and Y we have $S(Y,\,y) \geq S(X,\,y)$.*

This lemma follows from the observation that for any convex function f on the interval $[0,1]$ we have that $f(x)$ is less than $(1-x)f(0)+xf(1)$ and so we have the following.

$$\mathrm{E}\left[e^{\beta Y_i}\right] \leq \mathrm{E}\left[(1-Y_i)+Y_i e^{\beta}\right] = (1-\mathrm{E}[X_i])+\mathrm{E}[X_i]\,e^{\beta} = \mathrm{E}\left[e^{\beta X_i}\right]$$

So we get that $\ln Z(Y,\,\beta) \leq \ln Z(X,\,\beta)$ and the lemma then follows from (2).

4. Proof of Lemma 3

We now turn to the proof of Lemma 3. Formula (7) is proved by showing that $\beta(x)$ is the optimal β in (2). More specifically, we first note the following simple relations for $\beta \in (\beta_{\min},\,\beta_{\max})$.

$$\begin{aligned}
\frac{d\ln Z(X,\,\beta)}{d\beta} &= \frac{\mathrm{E}\left[Xe^{\beta X}\right]}{Z(X,\,\beta)} = \mathrm{E}_\beta[X] && (19)\\
\frac{d^2\ln Z(X,\,\beta)}{d\beta^2} &= \frac{\mathrm{E}\left[X^2e^{\beta X}\right]Z(X,\,\beta)-\mathrm{E}\left[Xe^{\beta X}\right]^2}{Z^2(X,\,\beta)}\\
&= \mathrm{E}_\beta\left[X^2\right]-\mathrm{E}_\beta[X]^2\\
&= \sigma^2(X,\,\beta) && (20)
\end{aligned}$$

To optimize (2) we now differentiate $x\beta-\ln Z(X,\,\beta)$ with respect to β. By (19) this derivative is $x-\mathrm{E}_\beta[x]$. Setting the derivative to zero gives $x=\mathrm{E}_\beta[x]$ or, by definition, $\beta=\beta(x)$. To see that this is a minimum note that (20) implies that the second derivative is $\sigma^2(\beta)$ and hence non-negative. Equation (8) follows from (7) and (5). Equation (10) follows from (19), (20), and (6).

To derive (9) we consider derivatives of the entropy $S(\beta(x))$ with respect to the deviation x.

$$\begin{aligned}
S(\beta(x)) &= x\beta(x) - \ln Z(X,\, \beta(x)) \\
\frac{dS(\beta(x))}{dx} &= \beta(x) + x\frac{d\beta(x)}{dx} - \frac{d\ln Z(X,\, \beta)}{d\beta}\frac{d\beta(x)}{dx} \\
&= \beta(x) + x\frac{d\beta(x)}{dx} - \mathrm{E}_{\beta(x)}[X]\,\frac{d\beta(x)}{dx} \\
&= \beta(x) \qquad (21)
\end{aligned}$$

$$\begin{aligned}
\frac{d^2S(\beta(x))}{dx^2} &= \frac{d\beta(x)}{dx} \\
&= 1/\left(\frac{d\mathrm{E}_\beta[X]}{d\beta}\right) \\
&= \frac{1}{\sigma^2(X,\, \beta(x))} \qquad (22)
\end{aligned}$$

Finally, (9) follows from (21), (22) and (6).

5. Negative Association

The analysis of the missing mass and histogram rule error involve sums of variables that are not independent. However, these variables are negatively associated—an increase in one variable is associated with decreases in the other variables. Formally, a set of real-valued random variables X_1, $\ldots$, X_n is negatively associated if for any two disjoint subsets I and J of the integers $\{1,\, \ldots,\, n\}$, and any two non-decreasing, or any two non-increasing, functions f from $R^{|I|}$ to R and g from $R^{|J|}$ to R we have the following.

$$\mathrm{E}[f(X_i,\, i \in I)g(X_j,\, j \in J)] \leq \mathrm{E}[f(X_i,\, i \in I)]\,\mathrm{E}[g(X_j,\, j \in J)]$$

Dubhashi and Ranjan (1998) give a survey of methods for establishing and using negative association. This section states some basic facts about negative association.

Lemma 5 *Let X_1, $\ldots$, X_n be any set of negatively associated variables. Let X'_1, $\ldots$, X'_n be independent shadow variables, i.e., independent variables such that X'_i is distributed identically to X_i. Let $X = \sum_i X_i$ and $X' = \sum_i X'_i$. For any set of negatively associated variables we have $S(X, x) \geq S(X', x)$.*

Proof

$$\begin{aligned}
Z(X,\, \beta) = \mathrm{E}\left[e^{\beta X}\right] &= \mathrm{E}\left[\Pi_i e^{\beta X_i}\right] \\
&\leq \Pi_i \mathrm{E}\left[e^{\beta X_i}\right] = \mathrm{E}\left[e^{\beta X'}\right] = Z(X', \beta)
\end{aligned}$$

The lemma now follows from the definition of S, i.e., Equation (2). ■

Lemma 6 *Let S be any sample of m items (ball throws) drawn IID from a fixed distribution on the integers (bins) $\{1, \ldots, V\}$. Let $\mathrm{c}(i)$ be the number of times integer i occurs in the sample. The variables $\mathrm{c}(1), \ldots, \mathrm{c}(V)$ are negatively associated.*

Lemma 7 *For any negatively associated variables $X_1, \ldots, X_n$, and any non-decreasing functions $f_1, \ldots, f_n$, we have that the quantities $f_1(X_1), \ldots, f_n(X_n)$ are negatively associated. This also holds if the functions f_i are non-increasing.*

Lemma 8 *Let $X_1, \ldots, X_n$ be a negatively associated set of variables. Let $Y_1 \ldots, Y_n$ be 0-1 (Bernoulli) variables such that Y_i is a stochastic function of X_i, i.e., $P(Y_i = 1 \mid X_1, \ldots, X_n, Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_n) = P(Y_i = 1 \mid X_i)$. If $P(Y_i = 1 \mid X_i)$ is a non-decreasing function of X_i then $Y_1, \ldots, Y_n$ are negatively associated. This also holds if $P(Y_i = 1 \mid X_i)$ is non-increasing.*

Proof

$$
\begin{aligned}
\mathrm{E}[f(Y_i,\ i \in I)g(Y_j,\ j \in J)] &= \mathrm{E}[\mathrm{E}[f(Y_i,\ i \in I)g(Y_j,\ j \in J) \mid X_1, \ldots, X_n]] \\
&= \mathrm{E}[\mathrm{E}[f(Y_i,\ i \in I)\ |X_i,\ i \in I]\,\mathrm{E}[g(Y_j,\ j \in J \mid X_j,\ j \in J)]] \\
&\leq \mathrm{E}[\mathrm{E}[f(Y_i,\ i \in I)\ |X_i,\ i \in I]]\,\mathrm{E}[\mathrm{E}[g(Y_j,\ j \in J \mid X_j,\ j \in J)]] \\
&= \mathrm{E}[f(Y_i,\ i \in I)]\,\mathrm{E}[g(Y_j,\ j \in J)]
\end{aligned}
$$

■

6. The Missing Mass: Inadequacy of Standard Inequalities

Suppose that we draw words (or any objects) independently from a fixed distribution over a countable (possibly infinite) set of words V. We let the probability of drawing word w be denoted as P_w. For a sample of m draws, the missing mass, denoted X, is the total probability mass of the items not occurring in the sample, i.e. $X = \sum_{w \notin S} P_w$. Let X_w be a Bernoulli variable which is 1 if word w does *not* occur in the sample and 0 otherwise. The missing mass can now be written as $X = \sum_w P_w X_w$. In the discussion below we let Q_w be $P(X_w = 1) \approx e^{-mP_w}$.

The variables X_w are monotonic functions of the word counts so by Lemmas 6 and 7 we have that the X_w are negatively associated. By Lemma 5 we can then assume that the variables X_w are independent. McAllester and Schapire (2000) prove that $S(X, \varepsilon) \geq m\varepsilon^2/3$ independent of the size of V or the distribution on V. Here we review why this result does not follow from standard inequalities for independent sums.

Hoeffding's inequality (Hoeffding, 1963) yields the following.

$$S(X, \mathrm{E}[X] + \varepsilon) \geq \frac{2\varepsilon^2}{\sum_w P_w^2}$$

In the missing mass application we have that $\sum_{i=1}^{V} P_w^2$ can be $\Omega(1)$ and so Hoeffding's inequality yields a bound that is $O(1)$ rather than the required $\Omega(m)$.

We now consider the Angluin-Valiant bound (13). Let $P_{\max} = \max_w P_w$. Define $Y = X/P_{\max}$. Lemma 4 implies that the Angluin-Valiant bound (13) generalizes to sums of independent variables bounded to $[0,1]$. Applying this generalization to Y yields the following for $\varepsilon \geq 0$.

$$\begin{aligned} S(X,\, \mathrm{E}_0[X]-\varepsilon) &= S(Y,\, \mathrm{E}_0[Y]-\frac{\varepsilon}{P_{\max}}) \\ &\leq \frac{\varepsilon^2}{2P_{\max}^2\mathrm{E}_0[Y]} \\ &= \frac{\varepsilon^2}{2P_{\max}\mathrm{E}_0[X]} \end{aligned} \tag{23}$$

To demonstrate the inadequacy of (23) take $|V|$ to be $m+1$, take P_u to be $1/2$ where u is a distinguished high probability word and take $P_w = 1/(2m)$ for $w \neq u$. We then have that $P_{\max} = 1/2$ and $\mathrm{E}_0[X] = \sum_{i=1}^{V} P_w Q_w \approx 1/(4e)$. In this case (23) is $O(1)$ rather than $\Omega(m)$.

An important observation for the missing mass is that $\sigma^2(X)$ is $O(1/m)$. In particular we have the following where we use negative association and $e^{-x} \leq 1/(ex)$ for $x \geq 0$.

$$\begin{aligned} \sigma^2(X) &\leq \sum_w P_w^2 Q_w(1-Q_w) \leq \sum_w P_w^2 Q_w \\ &\leq \sum_w P_w^2 e^{-mP_w} \leq \sum_w P_w/(em) \\ &\leq 1/(em) \end{aligned}$$

Since $\sigma^2(X)$ is $O(1/m)$ we might naturally hope to use Bernstein's inequality. The general form of Bernstein's inequality states that for $Y = \sum_w Y_w$ with Y_w independent, with zero mean, and with $Y_w \leq c$, we have the following.

$$S(X \geq \mathrm{E}[X]+\varepsilon) \geq \frac{\varepsilon^2}{2\sigma^2 + \frac{2}{3}c\varepsilon}$$

For the lower deviation of the missing mass we can take $Y_w = P_w(Q_w - X_w)$ in which case c can be taken to be $\max_w P_w Q_w \leq 1/(em)$. So Bernstein's inequality handles the downward deviation of the missing mass with a slightly weaker constant than that derived from (9) in Section 7. For the upward deviation of the missing mass, however, we take $Y_w = P_w(X_w - Q_w)$ in which case we need to take $c \geq \max_w P_w(1-Q_w)$. In this case the same example that defeats the Srivistav-Stranger bound defeats Bernstein's inequality—although $\sigma^2 \leq 1/(em)$ we have that $c\varepsilon$ can be as large as $c = 1/4$. So we get a bound that is $O(1)$ rather than $\Omega(m)$.

The general form of Bennett's inequality states that under the same conditions as Bernstein's inequality we have the following.

$$S(X \geq \mathrm{E}[X]+\varepsilon) \geq \frac{\sigma^2}{c^2} H\left(\frac{c\varepsilon}{\sigma^2}\right)$$

Here $H(t) = \iint_1^{1+t} 1/x\, d^2x = (1+t)\ln(1+t) - t$. The same example again defeats this bound for upward deviations of the missing mass. Again, while σ^2 is $O(1/m)$, we have that c and ε can both be $\Omega(1)$. This implies that $t = c\varepsilon/\sigma^2$ can be $\Omega(m)$ in which case we have the following.

$$\frac{\sigma^2}{c^2} H\left(\frac{c\varepsilon}{\sigma^2}\right) \leq O\left(\frac{1}{m}(1+m)\ln(1+m)\right)$$

So we get a bound that is $O(\ln m)$ rather than $\Omega(m)$.

7. A First Missing Mass Result

Here we derive both upper and lower concentration inequalities for the missing mass from (9) and (10) respectively. In Section 8 the constants in the upper bound will be improved. For the downward deviation we can use the following corollary of (9).

Lemma 9 *Let $X = \sum_i b_i X_i$ where the X_i are independent Bernoulli variables and $b_i \geq 0$. Let Q_i be $\mathrm{E}[X_i]$. For $\varepsilon \geq 0$ we have the following.*

$$S(X, \mathrm{E}[X] - \varepsilon) \quad \geq \quad \frac{\varepsilon^2}{2\sum_{i=1}^{V} Q_i b_i^2} \tag{24}$$

Formula (24) follows from (9) and the following which holds for $\beta \leq 0$.

$$\begin{aligned} \sigma^2(X,\, \beta) &= \sum_i Q_i(\beta)(1 - Q_i(\beta))b_i^2 \\ &\leq \sum_i Q_i(\beta) b_i^2 \\ &\leq \sum_i Q_i b_i^2 \end{aligned}$$

In the missing mass problem we have $Q_w \leq e^{-mP_w}$ and the argument in Section 6 showing that $\sigma^2(X) \leq 1/(em)$ also shows that $\sum_w P_w^2 Q_w \leq 1/(em)$. So we have the following downward deviation result which has a better constant than we get from Bernstein's inequality.

Theorem 10 *For the missing mass X as defined in Section 6, and for $\varepsilon \geq 0$, we have the following.*

$$S(X,\, \mathrm{E}[X] - \varepsilon) \quad \geq \quad \frac{em\varepsilon^2}{2}$$

We now derive an upward-deviation missing mass bound. This derivation is a variant of the derivation by McAllester and Schapire (2000) but benefits from the general statement of (10). We first note the following corollary of (10).

Lemma 11 *Let X be any real-valued random variable and let $\beta_{\max} \geq 0$ and $\sigma_{\max} \geq 0$ be constants such that for $0 \leq \beta \leq \beta_{\max}$ we have $\sigma^2(X,\, \beta) \leq \sigma_{\max}^2$. For $0 \leq \varepsilon \leq \beta_{\max}\sigma_{\max}^2$ we have the following.*

$$S(X,\, \mathrm{E}_0[X] + \varepsilon) \geq \frac{\varepsilon^2}{2\sigma_{\max}^2}$$

To see this note that by (10), for $0 \leq \beta \leq \beta_{\max}$ we have the following.

$$\ln Z(X,\, \beta) \leq \mathrm{E}_0[X]\beta + \frac{1}{2}\sigma_{\max}^2\beta^2$$

Inserting this into (2) and setting β equal to $\varepsilon/\sigma_{\max}^2 \leq \beta_{\max}$ proves the lemma.

For the missing mass problem we have $\mathrm{E}[X] \geq 0$ and Equation (10) implies that for $\beta \geq 0$ we have $Z \geq 1$. This implies that $P_\beta(X_w = 1) \leq Q_w e^{P_w \beta}$. This implies the following.

$$\sigma^2(X,\, \beta) \leq \sum_w P_w^2 P_\beta(X_w = 1) \leq \sum_w P_w^2 e^{-(m-\beta)P_w}$$

We now consider a constant $\beta_{\max}$ in the interval $[0,m]$. For any β in $[0,\beta_{\max}]$ we now have the following where we use $e^{-x} \leq 1/(ex)$ for $x \geq 0$.

$$\sigma^2(X,\ \beta) \quad \leq \quad \sum_w \frac{P_w}{e(m-\beta_{\max})} = \frac{1}{e(m-\beta_{\max})}$$

So for a given $\beta_{\max} \leq m$ we can take $\sigma^2_{\max} = 1/[e(m-\beta_{\max})]$. We can now apply Lemma 11 for $\varepsilon \leq \beta_{\max}/[e(m-\beta_{\max})]$. Solving for $\beta_{\max}$ as a function of ε gives the following.

$$\beta_{\max} = \frac{e\varepsilon m}{1+e\varepsilon}$$

Note that this satisfies $\beta_{\max} \leq m$. Solving for $\sigma^2_{\max}$ as a function of ε we get the following.

$$\sigma^2_{\max} = \frac{1+e\varepsilon}{em}$$

Applying Lemma 11 now yields the following.

Theorem 12 *For the missing mass X as defined in Section 6 we have the following for* $\varepsilon \geq 0$*.*

$$S(X,\ \mathrm{E}[X]+\varepsilon) \quad \geq \quad \frac{em\varepsilon^2}{2(1+e\varepsilon)}$$

For $0 \leq \varepsilon \leq 1$ *this gives the following.*

$$S(X,\ \mathrm{E}[X]+\varepsilon) \geq \frac{1}{3}m\varepsilon^2$$

8. A Second Solution

Now we give a concentration inequality for the upward-deviation of the missing mass that is not based on (9) or (10). Rather it is based on the following lemma of Kearns and Saul (1998).

Lemma 13 (Kearns and Saul) *For a Bernoulli variable Y we have the following where Q is* $P(Y=1)$*.*

$$\sup_\beta \frac{\ln Z(bY,\ \beta) - \mathrm{E}_0[bY]\beta}{\beta^2} \quad = \quad \frac{(1-2Q)b^2}{4\ln\frac{1-Q}{Q}} \tag{25}$$

$$\ln Z(bY,\ \beta) \quad \leq \quad \mathrm{E}_0[bY]\beta + \frac{(1-2Q)b^2}{4\ln\frac{1-Q}{Q}}\beta^2 \tag{26}$$

$$\leq \quad \mathrm{E}_0[bY]\beta + \frac{b^2}{4\ln\frac{1}{Q}}\beta^2 \tag{27}$$

We now make use of the following lemma whose proof is similar to the proof of Lemma 11. This lemma will also be important in Section 9.

Lemma 14 *If c is such that for all* β *we have* $\ln Z(X,\ \beta) \leq \mathrm{E}[X]\beta + c\beta^2$ *then (2) implies* $S(X,\ \mathrm{E}[X]+\varepsilon) \geq \varepsilon^2/(4c)$*.*

Formula (27) and observation 14 now immediately yield the following.

Lemma 15 *Let $X = \sum_{i=1}^{V} b_i X_i$ where the X_i are independent Bernoulli variables and $b_i \geq 0$. Let Q_i be $\mathrm{E}[X_i]$. For $\varepsilon \geq 0$ we have the following.*

$$S(X, \mathrm{E}[X] + \varepsilon) \quad \geq \quad \frac{\varepsilon^2}{\sum_{i=1}^{V} \frac{b_i^2}{\ln \frac{1}{Q_i}}}$$

Lemma 15 now immediately solves the upward-deviation of the missing mass problem. Let $X = \sum_w P_w X_w$ be the missing mass variable with $Q_w = P(X_w = 1) \leq e^{-mP_w}$. We now have the following.

$$\sum_w \frac{P_w^2}{\ln(1/Q_w)} \leq \sum_w \frac{P_w}{m} \leq \frac{1}{m}$$

Then, Lemma 15 yields the following.

Theorem 16 *For the missing mass X as defined in Section 6 we have the following.*

$$S(X,\, \mathrm{E}[X] + \varepsilon) \quad \geq \quad m\varepsilon^2$$

We have that theorem 16 is superior to theorem 12 for $\varepsilon \geq (1/2 - 1/e) \approx .07$.

9. Histogram Rule Error

Now we consider the problem of learning a histogram rule from an IID sample of pairs $\langle x, y\rangle \in \mathcal{X} \times \mathcal{Y}$ drawn from a fixed distribution D on such pairs. The problem is to find a rule h mapping $\mathcal{X}$ to the two-element set $\{0,1\}$ so as to minimize the expectation of the loss $l(h(x), y)$ where l is a given loss function from $\{0,1\} \times \mathcal{Y}$ to the interval $[0,1]$. In the classification setting one typically takes $\mathcal{Y}$ to be $\{0,1\}$. In the decision-theoretic setting y is the hidden state and can be arbitrarily complex and $l(h(x), y)$ is the cost of taking action $h(x)$ in the presence of hidden state y. In the general case (covering both settings) we assume only $h(x) \in \{0,1\}$ and $\ell(h(x), y) \in [0,1]$.

We are interested in histogram rules with respect to a fixed clustering. We assume a given cluster function C mapping $\mathcal{X}$ to the integers from 1 to k. We consider a sample S of m pairs drawn IID from a fixed distribution on $\mathcal{X} \times \mathcal{Y}$. For any cluster index j, we define S_j to be the subset of the sample consisting of pairs $\langle x, y\rangle$ such that $C(x) = j$. We define $\mathrm{c}(j)$ to be $|S_j|$. For any cluster index j and $w \in \{0,1\}$ we define $l_j(w)$ and $\hat{l}_j(w)$ as follows.

$$\hat{l}_j(w) = \frac{1}{\mathrm{c}(j)} \sum_{\langle x,y\rangle \in S_j} l(w, y), \qquad l_j(w) = \mathrm{E}_{\langle x,y\rangle \sim D \,|\, C(x)=j}\left[l(w, y)\right]$$

If $\mathrm{c}(j) = 0$ then we define $\hat{l}_j(w)$ to be 1. We now define the rule $\hat{h}$ and h^* from class index to labels as follows.

$$\hat{h}(j) = \operatorname*{argmin}_{w \in \{0,1\}} \hat{l}_j(w), \qquad h^*(j) = \operatorname*{argmin}_{w \in \{0,1\}} l_j(w)$$

Ties are broken stochastically with each outcome equally likely so that the rule $\hat{h}$ is a random variable only partially determined by the sample S. We are interested in the generalization loss of the empirical rule $\hat{h}$.

$$l(\hat{h}) = \mathrm{E}_{\langle x,y\rangle \sim D}\left[l(\hat{h}(C(x)), y)\right]$$

Theorem 17 *For $l(\hat{h})$ defined as above we have the following for positive* ε.

$$S\left(l(\hat{h}), \mathrm{E}\left[l(\hat{h})\right] - \varepsilon\right) \geq \frac{m\varepsilon^2}{7} \tag{28}$$

$$S\left(l(\hat{h}), \mathrm{E}\left[l(\hat{h})\right] + \varepsilon\right) \geq \frac{m\varepsilon^2}{9} \tag{29}$$

To prove this we need some additional terminology. For each class label j define P_j to be the probability over selecting a pair $\langle x, y\rangle$ that $C(x) = j$. Define L_j to be $l_j(1-h^*(j)) - l_j(h^*(j))$. In other words, L_j is the additional loss on class j when $\hat{h}$ assigns the wrong label to this class. Define the random variable X_j to be 1 if $\hat{h}(j) \neq h^*(j)$ and 0 otherwise. The variable X_j represents the statement that the empirical rule is "wrong" (non-optimal) on class j. We can now express the generalization loss of $\hat{h}$ as follows.

$$l(\hat{h}) = l(h^*) + \sum_i P_i L_i X_i$$

The variable X_j is a monotone stochastic function of the count $\mathrm{c}(j)$—the probability of error declines monotonically in the count of the class. By Lemma 8 we then have that the variables X_i are negatively associated so we can treat them as independent. To prove theorem 17 we start with an analysis of $P(X_j = 1)$.

Lemma 18

$$P(X_j = 1) \leq 3e^{-\frac{3}{16}mP_jL_j^2} \tag{30}$$

$$P(X_j = 1) \leq 3e^{-\frac{1}{2}(1-L_j)mP_jL_j^2} \tag{31}$$

Proof To prove this lemma we consider a threshold $y \leq mP_j$ and show the following.

$$P(X_j = 1) \leq P(\mathrm{c}(j) \leq y) + P(X_j = 1 \mid \mathrm{c}(j) \geq y) \tag{32}$$

$$P(\mathrm{c}(j) \leq y) \leq e^{-\left(mP_j - y\right)^2/(2mP_j)} \tag{33}$$

$$P(X_j = 1 \mid \mathrm{c}(j) \geq y) \leq 2e^{-2y\left(\frac{L_j}{2}\right)^2} \tag{34}$$

Formula (33) follows form the Angluin-Valiant bound. To prove (34) we note that if $X_j = 1$ then either $\hat{l}_j(h^*(j)) \geq l_j(h^*(j)) + L_j/2$ or $\hat{l}_j(1-h^*(j)) \leq l_j(1-h^*(j)) - L_j/2$. By a combination of Hoeffding's inequality and the union bound we have that the probability that one of these two conditions holds is bounded by the left hand side of (34). Setting y to $\frac{3}{8}mP_j$ yields the following which implies (30).

$$P(X_j = 1) \leq e^{-\frac{25}{128}mP_j} + 2e^{-\frac{3}{16}mP_jL_j^2}$$

Setting y to $(1-L_i)mP$ gives the following which implies (31).

$$P(X_j = 1) \leq e^{-\frac{1}{2}mP_jL_j^2} + 2e^{-\frac{1}{2}(1-L_j)mP_jL_j^2}$$

■

We now prove (28) using (30) and (10). Since X_i is bounded to the interval $[0,1]$ we have the following.

$$\sigma^2(P_iL_iX_i,\ \beta) \le \frac{1}{4}P_i^2L_i^2 \tag{35}$$

For $x \le \mathrm{E}[X]$ we have $\beta(x) \le 0$ and for $\beta \le 0$ we have the following.

$$\begin{aligned}
\sigma^2(P_iL_iX_i,\ \beta) &= P_i^2L_i^2P_\beta(X_i=1)(1-P_\beta(X_i=1)) \\
&\le P_i^2L_i^2P_\beta(X_i=1) \\
&\le P_i^2L_i^2P_0(X_i=1) \\
&\le P_i^2L_i^2 3e^{-\frac{3}{16}mP_iL_i^2}
\end{aligned} \tag{36}$$

Now let $\alpha = (16/3)\ln 12$. For i satisfying $mP_iL_i^2 \le \alpha$ we use (35) and for i satisfying $mP_iL_i^2 > \alpha$ we use (36). This gives the following where in deriving (37) we use the fact that xe^{-kx} is a monotonically decreasing function of x for $x > 1/k$.

$$\begin{aligned}
\ln Z(X,\ \beta) &\le \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_{mP_iL_i^2\le\alpha}\frac{P_i}{m}\left(\frac{mP_iL_i^2}{4}\right) + \sum_{mP_iL_i^2>\alpha}\frac{P_i}{m}\left(mP_iL_i^2 3e^{-\frac{3}{16}mP_iL_i^2}\right)\right)\beta^2 \\
&\le \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_{mP_iL_i^2\le\alpha}\frac{P_i}{m}\left(\frac{\alpha}{4}\right) + \sum_{mP_iL_i^2>\alpha}\frac{P_i}{m}\left(3\alpha e^{-\frac{3}{16}\alpha}\right)\right)\beta^2 && (37) \\
&= \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_i\frac{P_i}{m}\left(\frac{\alpha}{4}\right)\right)\beta^2 \\
&= \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\frac{\alpha}{4m}\right)\beta^2 && (38)
\end{aligned}$$

Formula (28) now follows from (38) and a variant of observation 14. The proof of (29) is similar. Let $\gamma = 16(2+\ln 3)/3$. For i satisfying $mP_iL_i^2 \le \gamma$ we use (35). For i satisfying $mP_iL_i^2 > \gamma$ we use (27) which yields the following.

$$\ln Z(P_iL_iX_i,\ \beta) \le \mathrm{E}_0[X]\beta + \left(\frac{P_i^2L_i^2}{4\left(\frac{3}{16}mP_iL_i^2 - \ln 3\right)}\right)\beta^2 \tag{39}$$

Combining (35) and (39) yields the following.

$$\begin{aligned}
\ln Z(X,\ \beta) &\le \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_{mP_iL_i^2\le\gamma}\frac{P_i}{m}\left(\frac{mP_iL_i^2}{4}\right) + \sum_{mP_iL_i^2>\gamma}\frac{P_i}{m}\left(\frac{mP_iL_i^2}{2\left(\frac{3}{16}mP_iL_i^2-\ln 3\right)}\right)\right)\beta^2 \\
&= \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_{mP_iL_i^2\le\gamma}\frac{P_i}{m}\left(\frac{mP_iL_i^2}{4}\right) + \sum_{mP_iL_i^2>\gamma}\frac{P_i}{m}\left(\frac{1}{2\left(\frac{3}{16}-\frac{\ln 3}{mP_iL_i^2}\right)}\right)\right)\beta^2 \\
&\le \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_{mP_iL_i^2\le\gamma}\frac{P_i}{m}\left(\frac{\gamma}{4}\right) + \sum_{mP_iL_i^2>\gamma}\frac{P_i}{m}\left(\frac{1}{2\left(\frac{3}{16}-\frac{\ln 3}{\gamma}\right)}\right)\right)\beta^2
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_{mP_iL_i^2 \leq \gamma} \frac{P_i}{m}\left(\frac{\gamma}{4}\right) + \sum_{mP_iL_i^2 > \gamma} \frac{P_i}{m}\left(\frac{\gamma}{2\left(\frac{3}{16}\gamma - \ln 3\right)}\right)\right)\beta^2 \\
&= \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\sum_i \frac{P_i}{m}\left(\frac{\gamma}{4}\right)\right)\beta^2 \\
&= \mathrm{E}_0[X]\beta + \frac{1}{2}\left(\frac{\gamma}{4m}\right)\beta^2 \qquad (40)
\end{aligned}
$$

Formula (29) now follows from (40) and Lemma 14.

10. Conclusions and Future Work

We have given concentration inequalities for the missing mass and for histogram rule error. In proving these results we have found it convenient to use a Gibbs-variance lemma relating Chernoff entropy to the variance of the Gibbs distribution for an arbitrary real valued random variable. There is a clear mismatch between the generality of the Gibbs-variance lemma, which applies to an arbitrary real valued random variable (an arbitrary configuration function) and the uses of this lemma here which are restricted to the analysis of independent sums. But it is not immediately obvious how to use the Gibbs-variance lemma in more complex settings such as those addressed by Talagrand's inequality or information-theoretic methods (McDiarmid, 1998, gives an overview of these methods). It seems likely that many quantities, possibly the cross-entropy of an n-gram language model, combine the extreme heterogeneity of the missing mass problem with coupling (non-independence) that defeats analysis by negative association. To prove concentration for such quantities it seems likely that sophisticated methods for handling difficult non-independence will need to be extended to handle extreme heterogeneity.

The concentration inequality for histogram rule error seems relevant in bounding the generalization error of decision trees. Fix a scheme for coding decision trees with finite bit strings. A use of Hoeffding's inequality, the Kraft inequality for codings, and the union bound can be used to show that, with probability at least $1-\delta$ over the choice of a training sample of size m, we have the following where $\ell(T)$ is the generalization of error of the tree T, $\hat{\ell}(T)$ is the error rate of T on the training data, and $|T|$ is the number of bits it takes to code T (McAllester, 1998).

$$
\forall T \;\; \ell(T) \leq \hat{\ell}(T) + \sqrt{\frac{(\ln 2)|T| + \ln(1/\delta)}{2m}} \qquad (41)
$$

Related results are given by Lugosi and Nobel (1996), Kearns and Mansour (1998), Langford and Blum (1999), and McAllester and Mansour (2000). A decision tree can be viewed as composed of two parts—a tree structure and the leaf labels. The tree structure defines a clustering—one cluster for each leaf—and the leaf labels specify a label for each cluster. If we allow a compact representation of tree structure then the number of bits needed to specify the structure can be small compared to the number of bits needed to specify the leaf labels. For example, suppose that we want to classify items with d binary-valued features. We can specify a tree structure by specifying a sequence of k features to test. In this case the number of bits needed to specify the tree structure is $O(k \ln d)$ but the number of bits needed for the leaf labels equals the number of leaves which is 2^k. More flexible methods of describing tree structure compactly are of course possible. We would like a version of (41) that is penalized by the bit length of the compactly described tree structure rather than the structure complexity plus the number of leaves.

Now consider a tree structure T and let $T(S)$ be the histogram rule defined by the clustering defined by T and the sample S. The concentration inequality for histogram rules implies that for a fixed T we have that the error rate $\ell(T(S))$ is near its expectation with high confidence. Suppose that we could find an error estimator $\hat{\ell}(T,S)$ satisfying the following two properties.

$$\mathrm{E}\left[\hat{\ell}(T,\, S)\right] \quad \geq \quad \mathrm{E}\left[\ell(T(S))\right] \tag{42}$$

$$S(\hat{\ell}(T,S),\, \hat{\ell}(T,S)+\varepsilon) \quad \geq \quad \Omega(m\varepsilon^2) \tag{43}$$

For any such estimator $\hat{\ell}(T,\, S)$ we have the following with probability at least $1-\delta$ over the choice of the sample where T ranges over tree structures rather than full trees.

$$\forall T \;\; \ell(T(S)) \leq \hat{\ell}(T,\, S) + O\left(\sqrt{\frac{|T| + \ln(1/\delta)}{m}}\right) \tag{44}$$

Formula (44) significantly improves on (41) in the case where the tree structures can be compactly described and the estimator $\hat{\ell}(T,\, S)$ has an expectation near that of $\ell(T(S))$. Formula (44) provides a foundation for data-dependent selection of tree structures—for a given sample one searches for a tree structure minimizing the bound on generalization error.

The difficulty in using (44) is finding an appropriate estimator $\hat{\ell}(T,\, S)$. The empirical error $\hat{\ell}(T(S)) = \mathrm{E}_{\langle x,y\rangle\sim S}\left[\ell(T(S)(x),y)\right]$ has the problem that the labels can overfit—consider the case where each cluster typically has only one data point. The empirical error of $T(S)$ fails to satisfy (42). The leave-one-estimator of $\ell(T(S))$ satisfies (42) but fails to satisfy (43). An interesting candidate is the Laplace error estimator defined as follows, for some constant a.

$$\begin{aligned}
\hat{\ell}(T,\, S) &= \sum_i \hat{P}_i \min(\hat{Q}_i, 1-\hat{Q}_i) \\
\hat{P}_i &= |S_i|/n \\
\hat{Q}_i &= \left(|\{\langle x,y\rangle \in S_i :\; y=1\}| + a\right)/(n+2a)
\end{aligned}$$

This error estimator satisfies the conditions of McDiarmid's theorem and hence satisfies (43). For sufficiently large values of the constant a it may also satisfy (42). Another approach to deriving a bound like (44) is to consider least squares regression rather than classification. For least squares regression the leave-one-out error estimate of $\ell(T(S))$ satisfies both (42) and (43). However, we have not yet proved a concentration inequality for $\ell(T(S))$ for least-squares regression. We leave the problem of derivation a concrete version of (44) for future work.

References

D. Angluin and L. Valiant. Fast probabilistic algorithms for Hamiltonian circuits. *Journal of Computing Systems Science*, 18:155–193, 1979.

S. Chen and J. Goodman. An empirical study of smoothing techniques for language modeling, August 1998. Technical report TR-10-98, Harvard University.

H. Chernoff. A measure of the asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

K. W. Church and W. A. Gale. A comparison of the enhanced Good-Turing and deleted estimation methods for estimating probabilities of English bigrams. *Computer Speech and Language*, 5: 19–54, 1991.

D. P. Dubhashi and D. Ranjan. Balls and bins: A study in negative dependence. *Random Structures and Algorithms*, 13(2):99–124, 1998.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(16):237–264, December 1953.

T. Hagerup and C. Rüb. A guided tour of Chernoff bounds. *Information Processing Letters*, 33: 305–309, 1989.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35(3): 400–401, March 1987.

M. Kearns and Y. Mansour. A fast, bottom-up decision tree pruning algorithm with near-optimal generalization. In *Proceedings of the 15th International Conference on Machine Learning*. Morgan Kaufmann, 1998.

M. Kearns and L. Saul. Large deviation methods for approximate probabilistic inference, with rates of convergence. In *Proceedings of Uncertainty in Artificial Intelligence 1998*, pages 311–319. Morgan Kaufmann, 1998.

S. Kutin. *Algorithmic Stability and Ensemble-Based Learning*. PhD thesis, University of Chicago, 2002.

J. Langford and A. Blum. Microchoice bounds and self-bounding learning algorithms. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 209–214, 1999.

G. Lugosi and A. Nobel. Consistency of data-driven histogram methods for density estimation and classification. *Annals of Statistics*, 24(2):687–706, 1996.

D. McAllester. Some PAC-Bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 230–234, 1998.

D. McAllester and Y. Mansour. Generalization bounds for decision trees. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 69–74, 2000.

D. McAllester and R. Schapire. On the convergence rate of Good-Turing estimators. In *Proceedings of the Thirteenth Annual Conference on Computational Learning Theory*, pages 1–6, 2000.

C. McDiarmid. Concentration. In M.Habib, C.McDiarmid, J.Ramirez-Alfonsin, and B.Reed, editors, *Probabilistic methods for algorithmic discrete mathematics*, pages 195–248. Springer, 1998.

L. E. Ortiz and L. Pack-Kaelbling. Sampling methods for action selection in influence diagrams. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence*, pages 378–385, 2000.