The 17th Annual Conference on Learning Theory, COLT 2004

Data Dependent Risk Bounds for Hierarchical Mixture of Experts Classifiers

Arik Azran and Ron Meir

Department of Electrical Engineering
Technion, Haifa 3200
Israel
(arik,rmeir)@(tx,ee).technion.ac.il

Abstract. The hierarchical mixture of experts architecture provides a flexible procedure for implementing classification algorithms. The classification is obtained by a recursive soft partition of the feature space in a data-driven fashion. Such a procedure enables *local classification* where several experts are used, each of which is assigned with the task of classification over some subspace of the feature space. In this work, we provide data-dependent generalization error bounds for this class of models, which lead to effective procedures for performing model selection. Tight bounds are particularly important here, because the model is highly parameterized. The theoretical results are complemented with numerical experiments based on a randomized algorithm, which mitigates the effects of local minima which plague other approaches such as the expectation-maximization algorithm.

1 Introduction

The mixture of experts (MoE) and hierarchical mixture of experts (HMoE) architectures, proposed in [10] and extensively studied in later work, is a flexible approach to constructing complex classifiers. In contrast to many other approaches, it is based on an adaptive soft partition of the feature space into regions, to each of which is assigned a 'simple' (e.g. generalized linear model (GLIM)) classifier. This approach should be contrasted with more standard approaches which construct a complex parameterization of a classifier over the full space, and attempt to learn its parameters.

In binary pattern classification one attempts to choose a soft classifier f from some class $\mathcal{F}$, in order to classify an *observation* $x \in \mathbb{R}^k$ into one of two classes $y \in \mathcal{Y} = \{-1, +1\}$ using $\mathrm{sgn}(f(x))$. In the case of the $0-1$ loss, the ideal classifier minimizes the *risk* $P_e(f) = \mathbf{P}\{\mathrm{sgn}(f(X)) \neq Y\} = \mathbf{P}\{Yf(x) \leq 0\}$. If $\mathrm{sgn}\,(\mathcal{F})$ consists of all possible mappings from $\mathbb{R}^k$ to $\mathcal{Y}$, then the ultimate best classifier is the Bayes classifier $f_B(X) = \mathrm{argmax}_{y\in\mathcal{Y}}\, \mathbf{P}\{Y = y|X\}$. In practical situations, the selection of a classifier is based on a sample $D_N = \{(X_n, Y_n) \in \mathcal{X} \times \mathcal{Y}\}_{n=1}^N$, where each pair is assumed to be drawn i.i.d. from an unknown distribution $P(X, Y)$.

In this paper we consider the class of hierarchical mixtures of experts classifiers [10], which is based on a soft adaptive partition of the input space, and a utilization of a small number of 'expert' classifiers in each domain. Such a procedure can be thought of, on the one hand, as extending standard approaches based on mixtures, and, on the other hand, providing a soft probabilistic extension of decision trees. This architecture has been successfully applied to regression, classification, control and time series analysis. It should be noted that since the HMoE architecture is highly parameterized, it is important to obtain tight error bounds, in order to prevent overfitting. Previous results attempting to establish bounds on the estimation error of the MoE system were based on the VC dimension [9] and covering number approaches [15]. Unfortunately, such approaches are too weak to be useful in any practical setting.

2 Preliminary results

Consider a soft classifier f, and the $0-1$ loss incurred by it, given by $I[yf(x) \leq 0]$, where $I[t \leq 0]$ is the indicator function of the event '$t \leq 0$'. While we attempt to minimize the expected value of the $0-1$ loss, it turns out to be inopportune to directly minimize functions based on this loss. First, the computational task is often intractable due to its non-smoothness. Second, minimizing the empirical $0-1$ loss may lead to severe overfitting. Many recent approaches are based on minimizing a smooth convex function $\phi(yf(x))$ which upper bounds the $0-1$ loss (e.g. [20, 12, 1]). Define the ϕ-risk, $E_\phi(f) = \mathbf{E}\left\{\phi(Yf(X))\right\}$, and denote the *empirical ϕ-risk* by $\hat{E}_\phi(f, D_N) = N^{-1}\sum_{n=1}^{N}\phi(y_n f(x_n))$. We assume that the loss function $\phi(t)$ satisfies $\phi(0) = 1$, $\phi(t)$ is Lipschitz with constant L_ϕ, $\phi_{\max} < \infty$ where $\phi_{\max} = \sup_{t\in\mathbb{R}}\phi(t)$ and $I[t \leq 0] \leq \phi(t)$ for all t. Using the ϕ-risk instead of the risk itself is motivated by several reasons. (i) Minimizing the ϕ-risk often leads asymptotically to the Bayes decision rule [20]. (ii) Rather tight upper bounds on the risk may be derived for finite sample sizes (e.g. [20, 12, 1]). (iii) Minimizing the empirical ϕ-risk instead of the empirical risk is computationally much simpler.

Data dependent error bounds are often derived using the Rademacher complexity. Let $\mathcal{F}$ be a class of real-valued functions with domain $\mathbb{R}^k$. The empirical Rademacher complexity is defined as

$$\hat{R}_N(\mathcal{F}) = \mathbf{E}_{\boldsymbol{\sigma}}\left\{\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{n=1}^{N}\sigma_n f(x_n)\right\} ,$$

where $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, ..., \sigma_N)$ is a random vector consisting of independently distributed binary random variables with $\mathbf{P}(\sigma_n = 1) = \mathbf{P}(\sigma_n = -1) = 1/2$. The Rademacher complexity is defined as the average over all possible training sequences, $R_N(\mathcal{F}) = \mathbf{E}_{D_N}\hat{R}_N(\mathcal{F})$.

The following Theorem, adapted from [2] and [16], will serve as our starting point.

Theorem 1. *For every $\delta \in (0,1)$ and positive integer N, with probability at least $1-\delta$ over training sequences of length N, every $f \in \mathcal{F}$ satisfies*

$$P_e(f) \le \hat{E}_\phi(f, D_N) + 2L_\phi \hat{R}_N(\mathcal{F}) + 3\phi_{\scriptscriptstyle max}\sqrt{\frac{\ln\frac{2}{\delta}}{2N}} \; .$$

This bound is proved in three steps. First McDiarmid's inequality [14] and a symmetrization argument [19] are used to bound $\sup_{f\in\mathcal{F}}(E_\phi(f) - \hat{E}_\phi(f, D_N))$ with $R_N(\phi \circ \mathcal{F})$, which is then bounded by $\hat{R}_N(\phi \circ \mathcal{F})$ using McDiarmid's inequality again. The claim is established by using the Lipschitz property of $\phi(\cdot)$ to bound $\hat{R}_N(\phi \circ \mathcal{F})$ with $L_\phi \hat{R}_N(\mathcal{F})$ (e.g. [11, 16]). In the sequel we upper bound $\hat{R}_N(\mathcal{F})$ for the case where $\mathcal{F}$ is the HMoE classifier.

Remark 1. The results of the Theorem can be tightened using the entropy method [4]. This leads to improved constants in the bounds, which are of particular significance when the sample size is small. We defer discussion of this issue to the full paper.

3 Mixture of Experts Classifiers

Consider initially the simple MoE architecture defined in Figure 1, and given mathematically by

$$f(x) = \sum_{m=1}^{M} a_m(w_m, x) h_m(v_m, x). \tag{1}$$

We interpret the functions h_m as *experts*, each of which 'operates' in regions of space for which the *gating functions* a_m are nonzero. Note that assuming a_m to be independent of x leads to a standard mixture. Such a classifier can be intuitively interpreted as implementing the principle of 'divide and conquer' where instead of solving one complicated problem (over the full space), we can do better by dividing it into several regions, defined through the gating functions a_m, and using 'simple' expert h_m in each region. It is clear that some restriction needs to be imposed on the gating functions and experts, since otherwise overfitting is imminent. We formalize the assumptions regarding the experts and gating functions below. These assumptions will later be weakened.

Definition 1 (Experts). *For each $1 \le m \le M$, let $\mathrm{V}^m_{\scriptscriptstyle max}$ be some nonnegative scalar and v_m a vector with k elements. Then, the m-th expert is given by a mapping $h_m(v_m, x)$ where $v_m \in V_m = \left\{v \in \mathbb{R}^k : \|v\| \le \mathrm{V}^m_{\scriptscriptstyle max}\right\}$. We define the collection of all functions $h_m(v_m, x)$ such that $v_m \in V_m$ as $\mathcal{H}_m$. To simplify the notation we define $\mathrm{V}_{\scriptscriptstyle max} = \sup_m \mathrm{V}^m_{\scriptscriptstyle max}$ and set $\mathcal{H} = \bigcup_{m=1}^M \mathcal{H}_m = \bigcup_{m=1}^M \left\{h_m(v_m, x),\ v_m \in V_m\right\}$.*

In the definitions below we write $\sup_{w_m, v_m}$ instead of $\sup_{w_m \in W_m, v_m \in V_m}$.

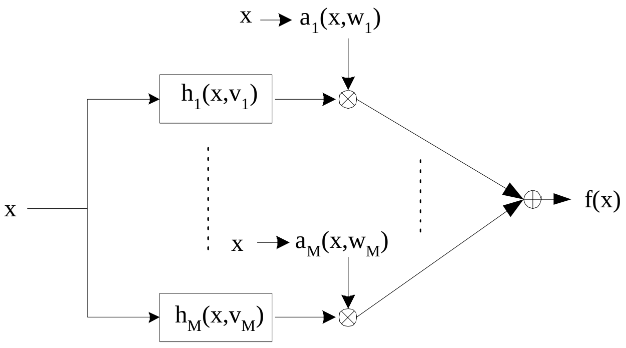

Fig. 1. MoE classifier with M experts.

Assumption 1 *The following assumptions, serving the purpose of regularization, are made for each* m, $1 \leq m \leq M$. *(i) To allow different types of experts, assume* $h_m(v_m, x) = h_m(\tau_m(v_m, x))$ *where* $\tau_m(v_m, x)$ *is some mapping such as* $v_m^\top x$ *or* $\|v_m - x\|$. *We assume that* $h_m(\tau_m(v_m, x))$ *is Lipschitz with constant* L_{h_m}, *i.e.* $|h_m(\tau_m(v_{m_1}, x)) - h_m(\tau_m(v_{m_2}, x))| \leq L_{h_m} |\tau_m(v_{m_1}, x) - \tau_m(v_{m_2}, x)|$. *(ii)* $|h_m(v_m, x)|$ *is bounded by some positive constant* $\mathcal{M}_{\mathcal{H}_m} < \infty$. *So, by defining* $\mathcal{M}_{\mathcal{H}} = \max_m \mathcal{M}_{\mathcal{H}_m}$ *we have that* $\sup_{m,v_m} |h_m(v_m, x)| \leq \mathcal{M}_{\mathcal{H}}$. *(iii) The experts are either symmetric (for regression) or antisymmetric (for classification) with respect to the parameters so that* $h_m(v_m, x) = \nu h_m(-v_m, x)$ *for some* $\nu \in \{\pm 1\}$.

Remark 2. Throughout our analysis we refer to x as a sample of the feature space. Yet, our results can be immediately extended to experts of the form $h_m(v_m, x) = h_m\left(v_m^\top \Phi_m(x)\right)$ where $\Phi_m(x)$ may be a high-dimensional nonlinear mapping as is used in kernel methods. Since our results are independent of the dimension of Φ_m, they can be used to obtain useful bounds for local mixtures of kernel classifiers. The use of such experts results in a powerful classifier that may select a different kernel in each region of the feature space.

The gating functions $a(\cdot, x)$ reflect the relative weights of each of the experts at a given point x. In the sequel we consider two main types of gating functions.

Definition 2 (Gating functions). *For each* $1 \leq m \leq M$, *let* W_{max}^m *be a nonnegative scalar and* w_m *a vector with* k *elements. Then, the* m*-th gating function is given by a mapping* $a_m(w_m, x)$ *where* $w_m \in W_m = \{w \in \mathbb{R}^k : \|w\| \leq \mathrm{W}_{max}^m\}$. *To simplify the notation we define* $\mathrm{W}_{max} = \sup_m \mathrm{W}_{max}^m$ *and set* $\mathcal{A} = \bigcup_{m=1}^M \mathcal{A}_m = \bigcup_{m=1}^M \{a_m(w_m, x) | w_m \in W_m\}$. *If* $a_m(w_m, x) = a_m(w_m^\top x)$ *we say that* $a_m(w_m, x)$ *is a* half-space gate, *and if* $a_m(w_m, x) = a_m\left(\|w_m - x\|^2/2\right)$ *we say that* $a_m(w_m, x)$ *is a* local gate.

Assumption 2 *The following assumptions are made for every m, $1 \leq m \leq M$. (i) $a_m(v_m, x)$ is Lipschitz with constant L_{a_m}, analogously to Assumption 1. We define $L_a = \max_m L_{a_m}$. (ii) $|a_m(v_m, x)|$ is bounded by some positive constant $\mathcal{M}_{\mathcal{A}_m} < \infty$. So, by defining $\mathcal{M}_{\mathcal{A}} = \max_m \mathcal{M}_{\mathcal{A}_m}$ we have $\sup_{m,w_m} |a_m(w_m, x)| \leq \mathcal{M}_{\mathcal{A}}$.*

In Section 6 we will remove some of the restrictions imposed on the parameters.

4 Risk bounds for mixture of experts classifiers

In this section we address the problem of bounding $\hat{R}_N(\mathcal{F})$ where $\mathcal{F}$ is the class of all MoE classifiers defined in section 3. We begin with the following Lemma, the proof of which can be found in the appendix.

Lemma 1. *Let $\mathcal{F}_m = \{a_m(w_m, x)h_m(v_m, x) | a_m(w_m, x) \in \mathcal{A}_m, h_m(v_m, x) \in \mathcal{H}_m\}$. Then, $\hat{R}_N(\mathcal{F}) = \sum_{m=1}^{M} \hat{R}_N(\mathcal{F}_m)$.*

Thus, it is suffices to bound $\hat{R}_N(\mathcal{F}_m)$, $m = 1, 2, \ldots, M$ in order to establish bounds for $\hat{R}_N(\mathcal{F})$. To do so, we use the following Lemma.

Lemma 2. *Let $\mathcal{G}_1, \mathcal{G}_2$ be two classes defined over some sets $\mathcal{X}_1, \mathcal{X}_2$ respectively, and define the class $\mathcal{G}_3$ as*

$$\mathcal{G}_3 = \{g : g(x_1, x_2) = g_1(x_1)g_2(x_2),\ g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}\ .$$

Assume further that at least one of the sets $\mathcal{X}_1$ or $\mathcal{X}_2$ is closed under negation and that every function in the class defined over this set is either symmetric or antisymmetric. Then,

$$\mathcal{Z}(\mathcal{G}_3) \leq \mathcal{M}_2 \mathcal{Z}(\mathcal{G}_1) + \mathcal{M}_1 \mathcal{Z}(\mathcal{G}_2)\ ,$$

where $\mathcal{Z}(\mathcal{G}_i) = \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g \in \mathcal{G}_i} \sum_{n=1}^{N} \sigma_n g(x_n) \right\}$ for $i = 1, 2, 3$ and $\mathcal{M}_i = \sup_{g_i \in \mathcal{G}_i, x_i \in \mathcal{X}_i} |g_i(x_i)|$ for $i = 1, 2$.

The proof of Lemma 2 is given in the Appendix. Note that a simpler derivation is possible using the identity $ab = (1/4)\left((a+b)^2 - (a-b)^2\right)$. However, this approach leads to looser bound. This lemma implies the following corollary.

Corollary 1. *For every $1 \leq m \leq M$ define $\mathcal{F}_m$ as in Lemma 1. Then,*

$$\hat{R}_N(\mathcal{F}_m) \leq \mathcal{M}_{\mathcal{H}_m} \hat{R}_N(\mathcal{A}_m) + \mathcal{M}_{\mathcal{A}_m} \hat{R}_N(\mathcal{H}_m) \qquad (m = 1, 2, \ldots, M)\ .$$

We emphasize that Corollary 1 is tight. To see that, set the gating function to be a constant. In this case $\hat{R}_N(\mathcal{A}_m) = 0$ and an equality is obtained by setting the gating variable to $\mathcal{M}_{\mathcal{A}_m}$. In the sequel we use the following basic result (see [11, 16] for a proof).

Lemma 3. *Assume ψ is Lipschitz with constant L_ψ and let $g : \mathbb{R}^k \times \mathcal{Y} \mapsto \mathbb{R}$ be some given function. Then, for every integer N*

$$\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{f \in \mathcal{F}} \sum_{n=1}^{N} \sigma_n \psi \left(g \left(y_n, f \left(x_n \right) \right) \right) \right\} \leq L_\psi \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{f \in \mathcal{F}} \sum_{n=1}^{N} \sigma_n g(y_n, f(x_n)) \right\} \ .$$

Remark 3. To minimize the technical burden, we assume the experts are generalized linear models (GLIM, see [13]), i.e. $\tau_m(v_m, x) = \tau_m(v_m^\top x)$ in Assumption 1. An extension to generalized radial basis functions (GRBF), i.e. $\tau_m(v_m, x) = \tau_m \left(\|v_m - x\| \right)$, is immediate using our analysis of local gating functions. Extensions to many other types can be achieved using similar technique.

Using the Lipschitz property of the class $\mathcal{H}_m$ along with Lemma 3 we get

$$\hat{R}_N(\mathcal{H}_m) \leq \frac{L_{h_m}}{N} \mathbf{E}_{\boldsymbol{\sigma}} \sup_{v_m} \left\{ v_m^\top \sum_{n=1}^{N} \sigma_n x_n \right\} \ .$$

By the Cauchy-Schwartz and Jensen inequalities we find that

$$\hat{R}_N(\mathcal{H}_m) \leq \frac{L_{h_m}}{N} \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \mathrm{V}^m_{\max} \left\| \sum_{n=1}^{N} \sigma_n x_n \right\| \right\} \leq \frac{L_{h_m} \mathrm{V}^m_{\max} \bar{x}}{\sqrt{N}}$$

where $\bar{x} = \sqrt{N^{-1} \sum_{j=1}^{k} \sum_{n=1}^{N} x_{nj}^2}$.

For the case of half-space gating functions we have $a(w, x) = a(w^\top x)$. In this case, analogous argumentation to the one used for the experts can be used to bound $\hat{R}_N(\mathcal{A})$. For the case of local gating functions we have $a(w, x) = a \left(\|w - x\|^2/2 \right)$. Similar arguments lead to the bound

$$\hat{R}_N(\mathcal{A}_m) \leq \frac{L_{a_m}}{\sqrt{N}} \left(\frac{(\mathrm{W}^m_{\max})^2}{\sqrt{8}} + \mathrm{W}^m_{\max} \bar{x} \right) .$$

We summarize our results in the following Theorem.

Theorem 2. *Let $\mathcal{F}$ be the class of mixture of experts classifiers with M GLIM experts. Assume that gates $1, 2, \ldots, M_1$ are local and that gates $M_1 + 1, \ldots, M$ are half-space where $0 \leq M_1 \leq M$. Then,*

$$\hat{R}_N(\mathcal{F}) \leq \frac{1}{\sqrt{N}} \left[\sum_{m=1}^{M_1} c_{1,m} (\mathrm{W}^m_{\mathit{max}})^2 + \sum_{m=1}^{M} c_{2,m} \mathrm{W}^m_{\mathit{max}} + \sum_{m=1}^{M} c_{3,m} \mathrm{V}^m_{\mathit{max}} \right]$$

where $c_{1,m} = \mathcal{M}_{\mathcal{H}_m} L_{a_m} / \sqrt{8}$, $c_{2,m} = \mathcal{M}_{\mathcal{H}_m} L_{a_m} \bar{x}$ and $c_{3,m} = \mathcal{M}_{\mathcal{A}_m} L_{h_m} \bar{x}$ for all $m = 1, 2, \ldots, M$.

5 Hierarchical Mixture of Experts

The MoE classifier is defined by a linear combination of M experts. An intuitive interpretation to the meaning of this combination is the division of the feature space into subspaces, in each of which the experts are combined using the weights a_m. The *Hierarchical* MoE takes this procedure one step further by recursively dividing the subspaces using a MoE classifier as the expert in each domain, as described in Figure 2.

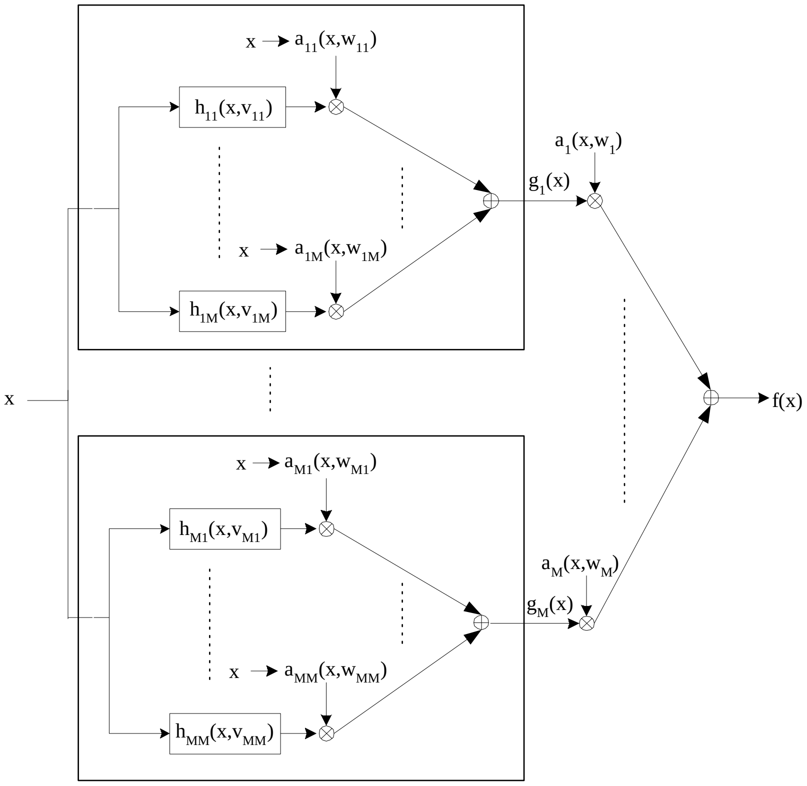

Fig. 2. Balanced 2-level HMoE classifier with M experts. Each expert in the first level is a mixture of M sub-experts.

In this section we expand the bound obtained for the MoE to the case of HMoE. We demonstrate the procedure for the case of balanced two-levels hierarchy with M experts (see Figure 2). It is easy to repeat the same procedure for any number of levels, whether the HMoE is balanced or not, using the same idea.

We begin by giving the mathematical description of the HMoE classifier. Let $f(x)$ be the output of the HMoE, and let $g_m(\theta_m, x)$ be the output of the m-th expert, $1 \leq m \leq M$. The parameter θ_m is comprised of all the parameters of the m-th first level expert, as will be detailed shortly. This is described by

$$f(x) = \sum_{m=1}^{M} a_m(w_m, x) g_m(\theta_m, x),$$

where $a_m(w_m, x)$ is the weight of the m-th expert in the first level $g_m(\theta_m, x)$, given by

$$g_m(\theta_m, x) = \sum_{j=1}^{M} a_{mj}(w_{mj}, x) h_{mj}(v_{mj}, x)$$

where $a_{mj}(w_{mj}, x)$ is the weight of $h_{mj}(v_{mj}, x)$, the j-th (sub-)expert in the m-th expert of the first level. By defining $\theta_{mj} = [w_{mj}, v_{mj}]$, we have that $\theta_m = [\theta_{m1}, \ldots, \theta_{mM}]$. We also define $w = [w_1, \ldots, w_M]$, the parameter vector of the gates of the first level and $\theta = [w, \theta_1, \ldots, \theta_M]$, the parameter vector of the HMoE.

Recall that we are seeking to bound the Rademacher complexity for the case of HMoE. First, we use the independence of the first level gating functions to show that

$$\hat{R}_N(\mathcal{F}) = \sum_{m=1}^{M} \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\theta_m, w_m} \frac{1}{N} \sum_{n=1}^{N} \sigma_n a_m(w_m, x_n) g_m(\theta_m, x_n) \right\} . \qquad (2)$$

So, our problem boils down to bounding the summands in (2). Notice that for every $m = 1, \ldots, M$ we have $\sup_{\theta_m}\{|g_m(\theta_m, x)|\} \leq M\mathcal{M}_{\mathcal{H}}\mathcal{M}_{\mathcal{A}}$. By defining $\mathcal{F}_m$ for the case of the 2-level HMoE analogously to the definition given at Lemma 1 for MoE, and using Corollary 1 recursively twice, it is easy to show that

$$\begin{aligned}
\hat{R}_N(\mathcal{F}_m) &\leq M\mathcal{M}_{\mathcal{H}}\mathcal{M}_{\mathcal{A}}\hat{R}_N(\mathcal{A}) + \mathcal{M}_{\mathcal{A}}\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\theta_m} \frac{1}{N} \sum_{n=1}^{N} \sigma_n \sum_{j=1}^{M} a_{mj}(w_{mj}, x_n) h_{mj}(v_{mj}, x_n) \right\} \\
&= M\mathcal{M}_{\mathcal{H}}\mathcal{M}_{\mathcal{A}}\hat{R}_N(\mathcal{A}) + \mathcal{M}_{\mathcal{A}} \sum_{j=1}^{M} \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\theta_{mj}} \frac{1}{N} \sum_{n=1}^{N} \sigma_n a_{mj}(w_{mj}, x_n) h_{mj}(v_{mj}, x_n) \right\} \\
&\leq M\mathcal{M}_{\mathcal{H}}\mathcal{M}_{\mathcal{A}}\hat{R}_N(\mathcal{A}) + M\mathcal{M}_{\mathcal{A}} \left(\mathcal{M}_{\mathcal{H}}\hat{R}_N(\mathcal{A}) + \mathcal{M}_{\mathcal{A}}\hat{R}_N(\mathcal{H}) \right) \\
&= M\mathcal{M}_{\mathcal{A}} \left[2\mathcal{M}_{\mathcal{H}}\hat{R}_N(\mathcal{A}) + \mathcal{M}_{\mathcal{A}}\hat{R}_N(\mathcal{H}) \right]
\end{aligned}$$

which, combined with Corollary 1 implies Theorem 3.

Theorem 3. *Let $\mathcal{F}$ be the class of balanced 2-level hierarchical mixture of experts classifiers with M experts in each division (see Figure 2). Then,*

$$\hat{R}_N(\mathcal{F}) \leq M^2\mathcal{M}_{\mathcal{A}} \left[2\mathcal{M}_{\mathcal{H}}\hat{R}_N(\mathcal{A}) + \mathcal{M}_{\mathcal{A}}\hat{R}_N(\mathcal{H}) \right] .$$

Notice that by choosing the constants more carefully, similar to Theorem 2, the bound in Theorem 3 can be tightened.

6 Fully data dependent bounds

So far, the feasible set for the parameters was determined by a ball with a predefined radius ($\mathrm{W}_{\max}$ for the gates or $\mathrm{V}_{\max}$ for the experts). This predefinition is problematic as it is difficult to know in advance how to set these parameters. Notice that given the number of experts M, these predefined parameters are the only elements in the bound that do not depend on the training sequence. In this section we eliminate the dependence on these preset parameters. Even though we give bounds for the case of MoE, the same technique can be easily harnessed to derive fully data dependent bounds for the case of HMoE.

The derivation is based on the technique used in [6]. The basic idea is to consider a grid of possible values for $\mathrm{W}^m_{\max}$ and $\mathrm{V}^m_{\max}$, for each of which Theorem 2 holds. Next, we assign a probability to each of these grid points and use a variant of the union bound to establish a bound that holds for every possible parameter.

Similarly to the definition of θ in section 5, we define for the MoE classifier $\theta = [\theta_1, \theta_2, \ldots, \theta_{2M}]$ where $\theta_m = w_m$ for all $m = 1, 2, \ldots, M$ and $\theta_m = v_m$ for all $m = M+1, M+2, \ldots, 2M$. The following result provides a data dependent risk bound with no preset system parameters, and can be proved using the methods described in [16].

Theorem 4. *Let the definitions and notation of Theorem 2 hold. Let q_0 be some positive number, and assume $\|\theta_m\| \geq q_0$ for every $m = 1, \ldots, 2M$. Then, with probability at least $1 - \delta$ over training sequences of length N, every function $f \in \mathcal{F}$ satisfies*

$$P_e(f) \leq \hat{E}_\phi(f, D_N) + \frac{2}{\sqrt{N}} \left[2 \sum_{m=1}^{M_1} c_{1,m} \|\theta_m\|^2 + \sum_{m=1}^{M} c_{2,m} \|\theta_m\| + \sum_{m=M+1}^{2M} c_{3,m} \|\theta_m\| \right] + 3\phi_{max} \sqrt{\ln \frac{2}{\delta} + 2 \sum_{m=1}^{2M} \ln \log_2 \frac{2\|\theta_m\|}{q_0}} \; .$$

Remark 4. Theorem 4 can be generalized to hold for all θ (without the restriction $\|\theta_m\| \geq q_0$), by using the proof method in [6],[16].

7 Algorithm and Numerical results

We demonstrate how the bound derived in Section 4 can be used to select the number of experts in the MoE model. We consider algorithms which attempt to minimize the empirical ϕ-loss $\hat{E}_\phi(f, D_N)$. It should be noted that previous methods for estimating the parameters of the MoE model were based on gradient methods for maximizing the likelihood or minimizing some risk function. Such approaches are prone to problems of local optima, which render standard gradient descent approaches of limited use. This problem also occurs for the EM algorithm discussed in [10]. Notice that even if $\phi(yf(x))$ is convex with respect to $yf(x)$, this doesn't necessarily imply that it is convex with respect to the parameters of $f(x)$. The deterministic annealing EM algorithm proposed in [17] attempts to address the local maxima problem, using a modified posterior distribution parameterized by a temperature like parameter. A modification of the EM algorithm, the split-and-merge EM algorithm proposed in [7], deals with certain types of local maxima involving an unbalanced usage of the experts over the feature space.

One possible solution to the problem of identifying the location of the global minimum of the loss is given by the *Cross-Entropy* algorithm (see [5] for a recent review, [18]). This algorithm, similarly to genetic algorithms, is based on the idea of randomly drawing samples from the parameter space and improving the way these samples are drawn from generation to generation. We observe that the algorithm below is applicable to finite dimensional problems.

To give an exact description of the algorithm used in our simulation we first introduce the following notation. We let the definition of θ from section 6 hold and denote by Θ the feasible set of values for θ. We also define a parameterized p.d.f. $\psi_\Theta(\theta;\xi)$ over Θ with ξ parameterizing the distribution.

To find a point that is likely to be in the neighborhood of the global minimum, we carry out Algorithm 1 (see box). Upon convergence, we use gradient methods with $\hat{\theta}_s^B$ (see box for definition) as the initial point to gain further accuracy in estimating the global minimum point. We denote by $\hat{\theta}^B$ the solution of the gradient minimization procedure and declare it as the final solution.

Simulation setup. We simulate a source generating data from a MoE classifier with 3 experts. The Bayes risk for this problem is 18.33%. We used a training sequence of length 300, for which we carried out Algorithm 1 followed by gradient search with respect to $\hat{E}_\phi(f, D_N)$, where $\phi(t) = 1 - \tanh(2t)$. Denoting by f_M^{CE} the classifier that was selected for each $M = 1, 2, \ldots, 5$, we denote by $\hat{E}_\phi(f_M^{CE}, D_N)$ the minimal empirical ϕ-risk obtained over the class. We evaluate the performance of each classifier by computing $\hat{P}_e(f_M^{CE}, D_{test})$ over a test sequence of 10^6 elements (D_{test}), drawn from the same source as the training sequence. This is the reported probability of error $P_e(f)$. Figure 1 describes these two measures computed over 400 different training sequences (the bars describe the standard deviation). The graph labelled as the 'complexity term' in Figure 1 is the sum of all terms on the right hand side of Theorem 2 with $\delta = 10^{-3}$, excluding $\hat{E}_\phi(f_M^{CE}, D_N)$. As for the CE parameters, we set $\psi_\Theta(.)$ to be the β distribution, $\hat{\xi}_0 = [1, 1]$ (corresponds to uniform distribution), $\rho_1 = 0.03$, $\rho_2 = 0.001$, $\rho_3 = 0.7$ and $T = 200$. The results are summarized in Figure 1.

A few observations are in place: (i) As one might expect, $\hat{E}_\phi(f_M^{CE}, D_N)$ is monotonically decreasing with respect to M. (ii) As expected, the complexity term is monotonically increasing with respect to M and (iii) $P_e(f)$ is the closest to the Bayes error (18.33%) when $M = 3$, which is the Bayes solution. We witness the phenomenon of underfitting for $M = 1, 2$ and overfitting for $M = 4, 5$, as predicted by the bound.

We also applied a variant of Algorithm 1, suitable for unbounded parameter feasible set (the details will be discussed in the full paper), to the real-world

The Cross-Entropy Algorithm.

Input: $\psi_\Theta(.)$ and $\phi(.)$.

Output: $\hat{\theta}_s^B$, a point in the neighborhood of the global minimum of $\hat{E}(f(\theta), D_N)$.

Algorithm :

1. Pick some $\hat{\xi}_0$ (a good selection will turn $\psi_\Theta(\theta; \hat{\xi}_0)$ into a uniform distribution over Θ). Set iteration counter $s = 1$, two positive integers d, T and three parameters $0 < \rho_1, \rho_2, \rho_3 < 1$.
2. Generate an ensemble $\theta_1, \ldots, \theta_L$ where $L = 2kMT$ (k is the dimension of the feature space and M is the number of experts, thus the dimension of Θ is $2kM$), drawn i.i.d according to $\psi_\Theta(\theta; \hat{\xi}_{s-1})$.
3. Calculate $\hat{E}_\phi(f, D_N)$ for each member of the ensemble. The Elite Sample (ES) comprises the $\lfloor \rho_1 L \rfloor$ parameters that received the lowest empirical ϕ-risk. Denote the parameters that are associated with the worst and the best $\hat{E}_\phi(f, D_N)$ in the ES as $\hat{\theta}_s^W$ and $\hat{\theta}_s^B$ respectively.
4. If for some $s \geq d$
$$\max_{s-d \leq i,j \leq s} (\hat{\theta}_i^W - \hat{\theta}_j^W) \leq \rho_2$$
stop (declare $\hat{\theta}_s^B$ as the solution). Otherwise, solve the maximum likelihood estimation problem, based on the ES, to estimate the parameters of ψ_Θ (notice that it is *not* a MLE for the original empirical risk minimization problem). Denoting the solution as $\hat{\xi}_{ML}$, compute $\hat{\xi}_{s+1} = (1 - \rho_3)\hat{\xi}_s + \rho_3 \hat{\xi}_{ML}$. Set $s = s + 1$ and return to 2.

Algorithm 1: The Cross-Entropy Algorithm for estimating the location of the global minimum of the empirical ϕ-risk.

data sets BUPA and PIMA [3]. We considered a MoE classifier with 1 to 4 linear experts, all with local gates. The results are compared with those of linear-SVM and RBF-SVM in Table 1.

Data set	MoE (2 experts)	Linear-SVM	RBF-SVM
BUPA	0.289 ± 0.050	0.320 ± 0.084	0.317 ± 0.048
PIMA	0.241 ± 0.056	0.244 ± 0.050	0.255 ± 0.067

Table 1. Real world data sets results. The results were computed using 7-fold cross-validation for BUPA and 10-fold cross-validation for PIMA. For each fold, the parameters of the classifiers were selected using cross-validation in the training sequence.

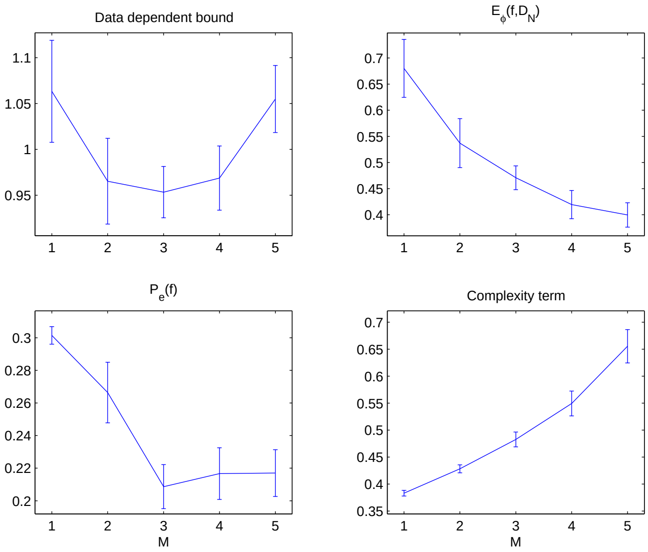

Fig. 3. A comparison between the data dependent bound of Theorem 2 and the true error, computed over 400 Monte Carlo iterations of different training sequences. The solid line describes the mean and the bars indicate the standard deviation over all training sequences. The two figures on the left demonstrates the applicability of the data dependent bound to the problem of model selection when one wishes to set the optimal number of experts. It can be observed that the optimal predicted value for M in this case is 3, which is the number of experts used to generate the data.

8 Discussion

We have considered the hierarchical mixture of experts architecture, and have established data dependent risk bounds for its performance. This class of architectures is very flexible and overly parameterized, and it is thus essential to establish bounds which do not depend on the number of parameters. Our bounds lead to very reasonable results on a toy problem. Also, the simulation results on real world problems are encouraging and motivate further research. Since the algorithmic issues are rather complicated for this architecture, it may be advantageous to consider some of the variational approaches proposed in recent years (e.g. [8]). We observe that the HMoE architecture can be viewed as a member of the large class of widely used graphical models (a.k.a. Bayesian networks). We expect that the techniques developed can be used to obtain tight risk bounds for these architectures as well.

Acknowledgment We are grateful to Dori Peleg for his assistance in applying the cross-entropy algorithm to our problem. The work of R.M. was partially supported by the Technion V.P.R. fund for the promotion of sponsored research.

Support from the Ollendorff center of the department of Electrical Engineering at the Technion is also acknowledged.

A Proofs of some of the theorems

Proof of Lemma 1 To simplify the notation, we write $\sup_{w,v}$ instead of $\sup_{w\in W, v\in V}$. Since, by definition, the set of parameters (w_i, v_i) is independent of (w_j, v_j) for any $1 \le i, j \le M$, $i \ne j$ we have

$$\begin{aligned}\hat{R}_N(\mathcal{F}) &= \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{w,v} \frac{1}{N} \sum_{n=1}^{N} \sigma_n \sum_{m=1}^{M} a_m(w_m, x_n) h_m(v_m, x_n) \right\} \\ &= \frac{1}{N} \sum_{m=1}^{M} \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{w_m, v_m} \sum_{n=1}^{N} \sigma_n a_m(w_m, x_n) h_m(v_m, x_n) \right\}.\end{aligned}$$

□

Proof of Lemma 2 First, we introduce the following Lemma

Lemma 4. *For any function $C(g_1, g_2, x)$, there exist $\nu \in \{\pm 1\}$ such that*

$$\begin{aligned}&\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1, g_2} \left(C(g_1, g_2, x) + \sigma g_1(x) g_2(x) \right) \right\} \\ \le &\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1, g_2} \left(C(g_1, \nu g_2, x) + \mathcal{M}_2 \sigma g_1(x) + \mathcal{M}_1 \sigma g_2(x) \right) \right\} \ .\end{aligned}$$

Proof. (of Lemma 4)

$$\begin{aligned}&\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1, g_2} \left(C(g_1, g_2, x) + \sigma g_1(x) g_2(x) \right) \right\} \\ &= \frac{1}{2} \sup_{g_1, g_2} \left(C(g_1, g_2, x) + g_1(x) g_2(x) \right) + \frac{1}{2} \sup_{g_1, g_2} \left(C(g_1, g_2, x) - g_1(x) g_2(x) \right) \\ &= \frac{1}{2} \sup_{g_1, g_2, \tilde{g}_1, \tilde{g}_2} \left(C(g_1, g_2, x) + g_1(x) g_2(x) + C(\tilde{g}_1, \tilde{g}_2, x) - \tilde{g}_1(x) \tilde{g}_2(x) \right) \\ &\overset{(a)}{=} \frac{1}{2} \sup_{g_1, g_2, \tilde{g}_1, \tilde{g}_2} \left(C(g_1, g_2, x) + C(\tilde{g}_1, \tilde{g}_2, x) + |g_1(x) g_2(x) - \tilde{g}_1(x) \tilde{g}_2(x)| \right) \\ &\overset{(b)}{\le} \frac{1}{2} \sup_{g_1, g_2, \tilde{g}_1, \tilde{g}_2} \left(C(g_1, g_2, x) + C(\tilde{g}_1, \tilde{g}_2, x) + \mathcal{M}_1 |g_2(x) - \tilde{g}_2(x)| + \mathcal{M}_2 |g_1(x) - \tilde{g}_1(x)| \right)\end{aligned} \tag{3}$$

where (a) is due to the symmetry of the expression over which the sepremum is taken and (b) is immediate, using the following inequality

$$\begin{aligned}|g_1(x) g_2(x) - \tilde{g}_1(x) \tilde{g}_2(x)| &= |g_1(x)(g_2(x) - \tilde{g}_2(x)) + \tilde{g}_2(x)(g_1(x) - \tilde{g}_1(x))| \\ &\le \mathcal{M}_1 |g_2(x) - \tilde{g}_2(x)| + \mathcal{M}_2 |g_1(x) - \tilde{g}_1(x)|.\end{aligned}$$

Next, we denote by $g_1^*, g_2^*, \tilde{g}_1^*, \tilde{g}_2^*$ the functions over which the supremum in (3) is achieved and address all cases of the signum of the terms inside the absolute values at (3).

case 1:$g_2^*(x) > \tilde{g}_2^*(x)$, $g_1^*(x) > \tilde{g}_1^*(x)$

$$\begin{aligned}
&\sup_{g_1,g_2,\tilde{g}_1,\tilde{g}_2} \{C(g_1,g_2,x) + C(\tilde{g}_1,\tilde{g}_2,x) + \mathcal{M}_1(g_2(x) - \tilde{g}_2(x)) + \mathcal{M}_2(g_1(x) - \tilde{g}_1(x))\}\\
&= \sup_{g_1,g_2} \{C(g_1,g_2,x) + \mathcal{M}_1 g_2(x) + \mathcal{M}_2 g_1(x)\} + \sup_{\tilde{g}_1,\tilde{g}_2} \{C(\tilde{g}_1,\tilde{g}_2,x) - \mathcal{M}_1\tilde{g}_2(x) - \mathcal{M}_2\tilde{g}_1(x)\}\\
&= 2\mathbf{E}_{\boldsymbol{\sigma}} \sup_{g_1,g_2} \{C(g_1,g_2,x) + \mathcal{M}_1\sigma g_2(x) + \mathcal{M}_2\sigma g_1(x)\}
\end{aligned}$$

case 2: $g_2^*(x) > \tilde{g}_2^*(x)$, $g_1^*(x) < \tilde{g}_1^*(x)$

$$\begin{aligned}
&\sup_{g_1,g_2,\tilde{g}_1,\tilde{g}_2} \{C(g_1,g_2,x) + C(\tilde{g}_1,\tilde{g}_2,x) + \mathcal{M}_1(g_2(x) - \tilde{g}_2(x)) + \mathcal{M}_2(\tilde{g}_1(x) - g_1(x))\}\\
&\overset{(a)}{=} \sup_{g_1,g_2,\tilde{g}_1,\tilde{g}_2} \{C(g_1,-g_2,x) + C(\tilde{g}_1,-\tilde{g}_2,x) + \mathcal{M}_1(\tilde{g}_2(x) - g_2(x)) + \mathcal{M}_2(\tilde{g}_1(x) - g_1(x))\}\\
&= 2\mathbf{E}_{\boldsymbol{\sigma}} \sup_{g_1,g_2} \{C(g_1,-g_2,x) + \mathcal{M}_1\sigma g_2(x) + \mathcal{M}_2\sigma g_1(x)\}
\end{aligned}$$

where (a) is due to the assumption that $\mathcal{G}_2$ is close under negation. Notice that the cases where $g_2^*(x) < \tilde{g}_2^*(x)$, $g_1^*(x) < \tilde{g}_1^*(x)$ and $g_2^*(x) < \tilde{g}_2^*(x)$, $g_1^*(x) > \tilde{g}_1^*(x)$ are analogous to cases 1 and 2 respectively. □

We can now provide the proof of Lemma 2. By using Lemma 4 recursively with a suitable definition of $C(g_1,g_2,x)$ in each iteration, we have for every $t = 1,\ldots,N+1$

$$\begin{aligned}
&\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1,g_2} \sum_{n=1}^{N} \sigma_n g_1(x_n) g_2(x_n) \right\}\\
&\leq \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1,g_2} \left(\sum_{n=t}^{N} \sigma_n g_1(x_n) g_2(x_n) + \mathcal{M}_2 \sum_{n=1}^{t-1} \sigma_n g_1(x_n) + \mathcal{M}_1 \sum_{n=1}^{t-1} \mathit{\Gamma}(n,t) \sigma_n g_2(x_n) \right) \right\}
\end{aligned}$$

where

$$\mathit{\Gamma}(n,t) = \begin{cases} \prod_{i=n}^{t-2} \nu_i & \text{if } n \leq t-2 \\ 1 & \text{if } n = t-1 \\ \text{not defined} & \text{otherwise} \end{cases} .$$

By setting $t = N+1$ we get

$$\begin{aligned}
&\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1,g_2} \sum_{n=1}^{N} \sigma_n g_1(x_n) g_2(x_n) \right\}\\
&\leq \mathcal{M}_2 \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_1} \sum_{n=1}^{N} \sigma_n g_1(x_n) \right\} + \mathcal{M}_1 \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{g_2} \sum_{n=1}^{N} \mathit{\Gamma}(n,N+1) \sigma_n g_2(x_n) \right\} .
\end{aligned}$$

Recall that $\nu_i \in \{\pm 1\}$ $\forall i$ and thus $\mathit{\Gamma}(n,N+1) \in \{\pm 1\}$ $\forall n$. So, by redefining $\sigma_n = \prod_{i=n}^{N-1} \nu_i \sigma_n$ $\forall n$ for the second term of the last inequality, we complete the proof of Theorem 2. □

References

1. Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification, and risk bounds. Technical Report 638, Department of Statistics, U.C. Berkeley, 2003.
2. P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
3. C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998.
4. S. Boucheron, G. Lugosi, and P. Massart. Concentration inequalities using the entropy method. *The Annals of Probability*, 31:1583–1614, 2003.
5. P.T. de Boer, D.P. Kroese, S. Mannor, and R.Y. Rubinstein. A tutorial on the cross-entropy method. *Annals of Operations Research*, 2004. To appear.
6. I. Desyatnikov and R. Meir. Data-dependent bounds for multi-category classification based on convex losses. In *Proc. of the sixteenth Annual Conference on Computational Learning Theory*, volume 2777 of *LNAI*. Springer, 2003.
7. Ghaharamani Z. Nakano R. Ueda N. Hinton, G.E. Smem algorithm for mixture models. *Neural Computation*, 12:2109–2128, 2000.
8. T. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods: Theory and Practice*, pages 129–159, Cambridge, MA, 2001. MIT Press.
9. W. Jiang. Complexity regularization via localized random penalties. *Neural Computation*, 12(6).
10. M.I. Jordan and R.A. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6(2):181–214, 1994.
11. M. Ledoux and M. Talgrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer Press, New York, 1991.
12. S. Mannor, R. Meir, and T. Zhang. Greedy algorithms for classification - consistency, convergence rates, and adaptivity. *Journal of Machine Learning Research*, 4:713–741, 2003.
13. P. McCullach and J. A. Nelder. *Generalized Linear Models.* CRC Press, 1989 (2nd edition).
14. C. McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, 1989.
15. R. Meir, R. El-Yaniv, and S. Ben-David. Localized boosting. In N. Cesa-Bianchi and S. Goldman, editors, *Proc. Thirteenth Annual Conference on Computaional Learning Theory*, pages 190–199. Morgan Kaufman, 2000.
16. R. Meir and T. Zhang. Generalization bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
17. R. Nakano and N. N. Ueda. Determinisic annealing em algorithm. *Neural Networks*, 11(2), 1998.
18. Rueven Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1:127–190, September 1999.
19. A.W. van der Vaart and J.A. Wellner. *Weak Convergence and EmpiricalProcesses.* Springer Verlag, New York, 1996.
20. T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1), 2004.