Constructing Multiclass Classifiers using Binary Classifiers Under Log-Loss

Assaf Ben-Yishai
Hebrew University of Jerusalem
assafbster@gmail.com

Or Ordentlich
Hebrew University of Jerusalem
or.ordentlich@mail.huji.ac.il

***Abstract*—The construction of multiclass classifiers from binary classifiers is studied in this paper, and performance is quantified by the regret, defined with respect to the Bayes optimal log-loss. We start by proving that the regret of the well known One vs. All (OVA) method is upper bounded by the sum of the regrets of its constituent binary classifiers. We then present a new method called Conditional OVA (COVA), and prove that its regret is given by the weighted sum of the regrets corresponding to the constituent binary classifiers. Lastly, we present a method termed Leveraged COVA (LCOVA), designated to reduce the regret of a multiclass classifier by breaking it down to independently optimized binary classifiers.**

I. Introduction

We consider the standard classification problem where $(X,Y) \sim P_{XY}$ are dependent random variables, drawn from a possibly unknown distribution P_{XY}, $Y \in \mathcal{Y} = \{0,...,K-1\}$ is the class label and $X \in \mathcal{X}$ is the observation. The goal is to come up with a classifier $f(X)$ that is close to Y with respect to some loss function. The most common loss function is the $0-1$ loss, and the corresponding classifier is designed such as to minimize the classification error probability $\Pr(f(X) \neq Y)$. In many cases, however, the observation X reveals some information on the label Y, but not enough to accurately predict the label. In such cases, a preferable approach is to design classifiers that output soft information, namely a conditional probability distribution for Y given X, rather than committing to a single value of Y. The common choice for a loss function measuring the quality of such a "soft classifier" is the logarithmic loss (log-loss)[1]–[6], which is the focus of this paper.

Since binary distributions are determined by a single parameter, a soft binary classifier amounts to a mapping from $\mathcal{X}$ to the interval $[0,1]$. Constructing a multiclass soft classifier, on the other hand, requires to choose a mapping from $\mathcal{X}$ to the $K-1$-dimensional simplex, which becomes a more complex task as K increases. It is therefore desirable to develop techniques for fusing multiple off-the-shelf binary classifiers (such as logistic regression, decision trees, support vector machines etc.) into a multiclass one. A good fusion method should have the property that if each of the binary classifiers are close to being optimal, then so is the resulting multiclass classifier. To that end, we define the regret of a classifier as the difference between the expected loss it attains and the loss attained by the optimal Bayes classifier. We then analyze the regret of fused multiclass classifiers in terms of the regrets of the underlying binary classifiers.

For the $0-1$ loss, many works have developed methods for constructing multiclass classifiers from binary ones [7]–[16] and have studied the dependence of the overall error probability on the error probabilities of the binary classifiers. Nevertheless, to the best of our knowledge, this is the first work to address this topic under log-loss.

Perhaps the simplest fusion method that comes to mind is the *One vs. All* method, where a binary soft classifier $\{p_i\}_{i=0}^{K-1}$ is constructed for each of the events $\{Y=i\}_{i=0}^{K-1}$, and those are merged to a distribution on $\mathcal{Y}$ simply by normalizing each p_i by $\sum_j p_j$. Our first main result is that the regret attained by this fusion method is upper bounded by the sum of the regrets of the binary classifiers. Next, we propose a novel merging method, dubbed *Conditional One vs. All* (COVA), inspired by the non-binary information-distilling qunatizer proposed in [17] and the non-binary channel upgrading algorithm in [18]. The COVA method is based on constructing binary soft classifiers for the events $\{Y=i\}$ conditioned on the event $\{Y \geq i\}$, for $i=0,\ldots,K-2$. Those classifiers are then naturally merged into a multiclass classifier, whose regret is proved to be *exactly* a weighted sum of the regrets of the underlying binary classifiers, where the weights are explicitly determined by P_Y.

In fact, any multiclass classifier induces $K-1$ probability assignments on the $K-1$ conditional events used by the COVA method. The regret of any such classifier is therefore the weighted sum of the regrets of those induced binary classifiers. Thus, if we can tweak a multiclass classifier in a way that decreases the log-loss of those induced binary classifiers, we are guaranteed to decrease the total multiclass log-loss. Based on this observation, we develop a method which we call *Leveraged COVA* (LCOVA) for improving the performance of parametric multiclass predictors. In particular, LCOVA gives rise to an improvement for the

978-1-5386-8209-8/21/$31.00 ©2021 IEEE

widely used Softmax multiclass classifier. Since Softmax forms the last layer in many deep neural net (DNN) architectures, the proposed LCOVA method may lead to an improverment in their respective performance.

II. Classifying Without Conditioning

For the sake of simplicity of exposition, we start by presenting our results without conditioning on the observed data X. The generalization to the conditional case is given in the following section. Let $Y \sim P$ be the class random variable supported on $\mathcal{Y} = \{0, \ldots, K-1\}$, and $K > 2$. We use $P(i)$ and p_i interchangeably to denote the class probability $\Pr(Y = i)$. Let Q denote a (possibly mismatched) probability assignment on $\mathcal{Y}$. We define the log-loss for predicting Y based on Q, while the actual underlying distribution is P by

$$L(P, Q) \triangleq \mathbb{E}_{y \sim P} \log \frac{1}{Q(y)}.$$

This quantity is also known as the *cross-entropy* of Q relative to P. It is well-known that $\min_Q L(P, Q)$ is attained by $Q = P$, and that for this choice $L(P, P) = H(Y)$, where $H(\cdot)$ denotes entropy. We denote the regret related to using Q instead of P by:

$$R(P, Q) \triangleq L(P, Q) - L(P, P) = D(P \,||\, Q),$$

where $D(P \,||\, Q)$ is the Kullback-Leibler divergence between P and Q. We shall use $R(P, Q)$ and $D(P \,||\, Q)$ interchangeably, where the first notation shall be used to state results, and the latter shall be used for the analysis.

In the sequel we use Bernoulli random variables and denote their respective properties (success probability, log-loss, regret etc.) using lowercase letters. Namely, let $U \sim \text{Ber}(p)$, (i.e. $U \in \{0, 1\}$, and $\Pr(U = 1) = p$), and let q be the parameter of a possibly mismatched distribution $\text{Ber}(q)$. The related (binary) log-loss is defined as $\ell(p, q) \triangleq p \log \frac{1}{q} + (1 - p) \log \frac{1}{1-q}$, and the related (binary) regret is defined as $r(p, q) \triangleq \ell(p, q) - \ell(p, p) = d(p \,||\, q)$, where $d(p \,||\, q)$ denotes the binary divergence and $\ell(p, p) = h(p)$ denotes the binary entropy of p. Let us define the following set of Bernoulli random variables

$$A_i = A_i(Y) \triangleq \mathbb{1}_{(Y=i)}, \quad i = 0, \ldots, K-1.$$

where $\mathbb{1}_{(\cdot)}$ is an indicator function, being equal to one if the condition is satisfied and zero otherwise. Trivially,

$$p_{A_i} \triangleq \Pr(A_i = 1) = P(i).$$

This identity implies that the set of success probabilities $\{p_{A_i}\}_{i=1}^{K-1}$ can provide the exact distribution of Y. We now present two methods for building estimators for P using Bernoulli random variables. The first is called *One vs. All (OVA)*, and is straightforward and widely used. The second, *Conditional OVA (COVA)*, is a novel contribution, and is slightly more complicated conceptually.

Definition 1 (One vs. All (OVA))**.** *Given a set of K estimates $\{q_{A_i}\}_{i=0}^{K-1}$, not all zero, of the respective probabilities $\{\Pr(A_i = 1)\}_{i=0}^{K-1}$, the OVA estimate of P is defined as*

$$Q^{\text{OVA}}(i) = \frac{q_{A_i}}{\sum_{j=0}^{K-1} q_{A_j}}, \quad i = 0, \ldots, K-1. \quad (1)$$

To motivate our suggested COVA method, we first express $\Pr(Y = i)$ in the following unconventional way

$$\begin{aligned} \Pr(Y = i) &= \Pr(Y = i, Y \geq i) \\ &= \Pr(A_i = 1 \mid Y \geq i) \Pr(Y \geq i). \end{aligned} \quad (2)$$

Noting that

$$\begin{aligned} \Pr(Y \geq i) &= \prod_{j=0}^{i-1} \Pr(Y \neq j \mid Y \neq 0, ..., Y \neq j-1) \\ &= \prod_{j=0}^{i-1} \Pr(A_j = 0 \mid Y \geq j), \end{aligned}$$

we get

$$P(i) = \Pr(A_i = 1 \mid Y \geq i) \prod_{j=0}^{i-1} \Pr(A_j = 0 \mid Y \geq j). \quad (3)$$

Denoting

$$p_{A_i}^{\text{cond}} \triangleq \Pr(A_i = 1 \mid Y \geq i) \quad (4)$$

and noticing that $\Pr(A_{K-1} = 1 \mid Y \geq K-1) = 1$, we can rewrite (3) as

$$P(i) = \begin{cases} p_{A_i}^{\text{cond}} \prod_{j=0}^{i-1} \left(1 - p_{A_j}^{\text{cond}}\right) & i < K-1 \\ \prod_{j=0}^{K-2} \left(1 - p_{A_j}^{\text{cond}}\right) & i = K-1 \end{cases} \quad (5)$$

Thus, given a set of possibly inaccurate estimates $\{q_i^{\text{cond}}\}_{i=0}^{K-1}$ for $\{p_i^{\text{cond}}\}_{i=0}^{K-1}$, we can plug them in (5) and obtain an estimate for P. This method summarized in the next definition.

Definition 2 (Conditional OVA (COVA))**.** *Given a set of $K-1$ estimates $\{q_i^{\text{cond}}\}_{i=0}^{K-2}$ for the respective conditional success probability $\{\Pr(A_i = 1 \mid Y \geq i)\}_{i=0}^{K-2}$, the COVA estimate to P is defined as*

$$Q^{\text{COVA}}(i) = \begin{cases} q_{A_i}^{\text{cond}} \prod_{j=0}^{i-1} \left(1 - q_{A_j}^{\text{cond}}\right) & i < K-1 \\ \prod_{j=0}^{K-2} \left(1 - q_{A_j}^{\text{cond}}\right) & i = K-1 \end{cases} \quad (6)$$

We note that Def. 2 can be regarded as a special case of hierarchical binary classification in the spirit of [12], [19]. This paper focuses on this special case for the sake of clarity of exposition. Generalizations of the definition (including permutation of the labels) and the respective regret appear in extended version [20].

A. The Regrets of OVA and COVA

We now bound the regrets $R(P, Q^{\mathrm{OVA}})$ of the OVA procedure and $R(P, Q^{\mathrm{COVA}})$ of the COVA method, in terms of the regrets of the involved binary classifiers, namely, the sets $\{r(p_i, q_{A_i})\}$, and $\{r(p^{\mathrm{cond}}_{A_i}, q^{\mathrm{cond}}_{A_i})\}$, respectively.

Lemma 1 (OVA regret)**.**

$$R(P, Q^{\mathrm{OVA}}) \leq \sum_{i=0}^{K-1} r(p_i, q_{A_i}) \tag{7}$$

Proof. We start by rewriting (1) as $Q^{\mathrm{OVA}}(i) = \frac{q_{A_i}}{\alpha K}$, for $i = 0, \ldots, K-1$, where $\alpha \triangleq \frac{\sum_{i=0}^{K-1} q_{A_i}}{K}$. Note that since $q_{A_i} \in [0, 1]$ for all i it is guaranteed that $\alpha \in (0, 1]$ (note that by Def. (1), $\{q_{A_i}\}$ are not all zero, so $\alpha > 0$). Recalling that the regrets can be written as divergences, the statement in (7) is equivalent to $\sum_{i=0}^{K-1} d(p_i \mid\mid q_{A_i}) - D(P \mid\mid Q^{\mathrm{OVA}}) \geq 0$. Expanding $D(P \mid\mid Q^{\mathrm{OVA}})$ yields

$$\begin{aligned} D(P \mid\mid Q^{\mathrm{OVA}}) &= \sum_{i=0}^{K-1} p_i \log \frac{p_i}{q_{A_i}/(\alpha K)} \\ &= \sum_{i=0}^{K-1} p_i \log \frac{p_i}{q_{A_i}} + \log(\alpha K), \end{aligned} \tag{8}$$

and expanding $\sum_{i=0}^{K-1} d(p_i \mid\mid q_{A_i})$ yields

$$\begin{aligned} &\sum_{i=0}^{K-1} d(p_i \mid\mid q_{A_i}) \\ &\quad = \sum_{i=0}^{K-1} \left(p_i \log \frac{p_i}{q_{A_i}} + (1-p_i) \log \frac{1-p_i}{1-q_{A_i}} \right). \end{aligned} \tag{9}$$

Subtracting (8) from (9) we obtain

$$\begin{aligned} &\sum_{i=0}^{K-1} d(p_i \mid\mid q_{A_i}) - D(P \mid\mid Q^{\mathrm{OVA}}) \\ &\quad = \sum_{i=0}^{K-1} (1-p_i) \log \frac{1-p_i}{1-q_{A_i}} - \log(\alpha K) = F_1 + F_2, \end{aligned}$$

where the last transition is by adding and subtracting the term $(K-1) \log \frac{K-1}{K(1-\alpha)}$ and defining

$$F_1 \triangleq \sum_{i=0}^{K-1} (1-p_i) \log \frac{1-p_i}{1-q_{A_i}} - (K-1) \log \frac{K-1}{K(1-\alpha)},$$

$$F_2 \triangleq (K-1) \log \frac{K-1}{K(1-\alpha)} - \log(\alpha K).$$

Note that since for all i, $p_i \leq 1$ and $q_{A_i} \leq 1$, it is guaranteed that $1 - p_i$ and $1 - q_{A_i}$ are non-negative (for $p_i = 1$ we use the convention that $0 \log 0 = 0$). Since $\sum_{i=0}^{K-1} p_i = 1$ and $\sum_{i=0}^{K-1} q_{A_i} = \alpha K$, $F_1 \geq 0$ holds due to the log-sum inequality [21]. Furthermore,

$$\begin{aligned} F_2 &= (K-1) \log \frac{K-1}{K(1-\alpha)} - \log(\alpha K) \\ &= K \left((1 - \tfrac{1}{K}) \log \frac{1 - \frac{1}{K}}{(1-\alpha)} + \tfrac{1}{K} \log \frac{\frac{1}{K}}{\alpha} \right) \\ &= K \cdot d\left(\tfrac{1}{K} \mid\mid \alpha \right) \geq 0, \end{aligned}$$

which concludes the proof. □

Lemma 2 (COVA regret)**.**

$$R(P, Q^{\mathrm{COVA}}) = \sum_{i=0}^{K-2} \Pr(Y \geq i) r(p^{\mathrm{cond}}_{A_i}, q^{\mathrm{cond}}_{A_i}).$$

Proof. Let us start by expanding the related losses. Starting with $L(P, Q^{\mathrm{COVA}})$ using (6) we obtain

$$\begin{aligned} &L(P, Q^{\mathrm{COVA}}) = \sum_{i=0}^{K-1} P(i) \log \frac{1}{Q^{\mathrm{COVA}}(i)} \\ &= \sum_{i=0}^{K-2} P(i) \log \frac{1}{q^{\mathrm{cond}}_{A_i} \prod_{j=0}^{i-1} (1 - q^{\mathrm{cond}}_{A_j})} \\ &\quad + P(K-1) \log \frac{1}{\prod_{j=0}^{K-1} (1 - q^{\mathrm{cond}}_{A_j})} \\ &= \sum_{i=0}^{K-2} \left(P(i) \log \frac{1}{q^{\mathrm{cond}}_{A_i}} + \Pr(Y > i) \log \frac{1}{1 - q^{\mathrm{cond}}_{A_i}} \right), \end{aligned}$$

where the last transition is by rearranging the sums. Substituting the definition of $p^{\mathrm{cond}}_{A_i}$ from (4) into (2) we have that $P(i) = p^{\mathrm{cond}}_{A_i} \Pr(Y \geq i)$ and similarly,

$$\begin{aligned} \Pr(Y > i) &= \Pr(Y > i, Y \geq i) \\ &= \Pr(Y > i \mid Y \geq i) \Pr(Y \geq i) \\ &= (1 - p^{\mathrm{cond}}_{A_i}) \Pr(Y \geq i) \end{aligned}$$

which renders

$$L(P, Q^{\mathrm{COVA}}) = \sum_{i=0}^{K-2} \Pr(Y \geq i) \ell(p^{\mathrm{cond}}_{A_i}, q^{\mathrm{cond}}_{A_i}). \tag{10}$$

To calculate $L(P, P)$ we repeat the same process, replacing Q^{COVA} by P, as expressed in (5) and obtain

$$L(P, P) = \sum_{i=0}^{K-2} \Pr(Y \geq i) \ell(p^{\mathrm{cond}}_{A_i}, p^{\mathrm{cond}}_{A_i}). \tag{11}$$

Subtracting (11) from (10) concludes the proof. □

III. Adding Conditioning on the Observations

Let us now consider a setup where one is interested in predicting Y based on some observation X, where $(X,Y) \sim P_{XY} = P_X \times P_{Y|X}$. The observation random variable X is supported on $\mathcal{X}$, where $\mathcal{X}$ is either some discrete alphabet or $\mathcal{X} = \mathbb{R}^d$. We denote the posterior probability of the label y given the observation x by

$$P_{Y|X=x}(y) \triangleq \Pr(Y = y \mid X = x).$$

The prediction here is "soft", that is, given the observation $X = x$ the goal is to provide a probability assignment $Q_{Y|X=x}$ which is close to $P_{Y|X=x}$ under log-loss. We would now like to construct $Q_{Y|X=x}$ for every value $x \in \mathcal{X}$, using a set of conditional Bernoulli success probabilities. Using the OVA methods, we denote

$$p_{A_i|X=x} \triangleq \Pr(A_i = 1 \mid X = x)$$

and denote its respective estimate by $q_{A_i|X=x}$. The OVA estimate now follows by adding the conditioning to (1) yielding for $0 \le i \le K-1$

$$Q^{\mathrm{OVA}}_{Y|X=x}(i) = \frac{q_{A_i|X=x}}{\sum_{j=0}^{K-1} q_{A_j|X=x}}. \tag{12}$$

Extending the COVA method to the conditional case is done similarly as follows. For every $x \in \mathcal{X}$ denote

$$p^{\mathrm{cond}}_{A_i|X=x} \triangleq \Pr(A_i = 1 \mid Y \ge i, X = x)$$

and its respective estimate by $q^{\mathrm{cond}}_{A_i|X=x}$. (6) is now extended to

$$Q^{\mathrm{COVA}}_{Y|X=x}(i) = \begin{cases} q^{\mathrm{cond}}_{A_i|X=x} \prod_{j=0}^{i-1} \left(1 - q^{\mathrm{cond}}_{A_j|X=x}\right) & i < K-1 \\ \prod_{j=0}^{K-2} \left(1 - q^{\mathrm{cond}}_{A_j|X=x}\right) & i = K-1 \end{cases} \tag{13}$$

We use $L(P_{Y|X}, Q_{Y|X} \mid P_X)$ to denote the expected conditional log-loss and define it as

$$L(P_{Y|X}, Q_{Y|X} \mid P_X) \triangleq \mathbb{E}_{x \sim P_X} L(P_{Y|X=x}, Q_{Y|X=x}).$$

The conditional regret is denoted and defined similarly:

$$R(P_{Y|X}, Q_{Y|X} \mid P_X) \triangleq \mathbb{E}_{x \sim P_X} R(P_{Y|X=x}, Q_{Y|X=x}),$$

which is also the standard definition of the conditional divergence $D(P_{Y|X} \mid\mid Q_{Y|X} \mid P_X)$. The Bernoulli counterparts are appropriately defined and denoted as follows:

$$\ell(p_{|X}, q_{|X} \mid P_X) \triangleq \mathbb{E}_{x \sim P_X} \ell(p_{|X=x}, q_{|X=x})$$
$$r(p_{|X}, q_{|X} \mid P_X) \triangleq \mathbb{E}_{x \sim P_X} r(p_{|X=x}, q_{|X=x})$$

Using these definitions, the following corollarries are obtained by taking the expectation over Lemma 1 and Lemma 2 respectively.

Corollary 1 (OVA conditional regret)**.**

$$R(P_{Y|X}, Q^{\mathrm{OVA}}_{Y|X} \mid P_X) \le \sum_{i=0}^{K-1} r(p_{i|X}, q_{A_i|X} \mid P_X).$$

Corollary 2 (COVA conditional regret)**.**

$$R(P_{Y|X}, Q^{\mathrm{COVA}}_{Y|X} \mid P_X) = \sum_{i=0}^{K-2} \Pr(Y \ge i) r(p^{\mathrm{cond}}_{A_i|X}, q^{\mathrm{cond}}_{A_i|X} \mid P_{X|Y \ge i})$$

A. Training the Binary Classifiers

In supervised learning under log-loss, one is given a training set of labeled samples $T \triangleq \{(x_i, y_i)\}_{n=1}^N$ drawn independently from an unknown distribution P_{XY}, and is required to output a conditional distribution $Q_{Y|X}$ for which the regret $R(P_{Y|X}, Q_{Y|X})$ is small. We are interested in a "black-box" reduction from the multiclass supervised learning problem to the binary case. To this end, assume we have access to an "off-the-shelf" binary classifier, (e.g., logistic regression, decision tree) which gets a training set with binary labels $\{(x_n, a_n)\}_{n=1}^N$, $x \in \mathcal{X}$, $a_n \in \{0,1\}$, as input, and constructs a probability assignment with small regret $q_{A|X=x}$ for every $x \in \mathcal{X}$ as output. We now describe how to train a low-regret multiclass classifier via this "black-box", using the OVA and the COVA methods.

In the OVA case, we build $Q^{\mathrm{OVA}}_{Y|X=x}$ using the set $\{q_{A_i|X=x}\}_{i=0}^{K-1}$ according to (12). Every $q_{A_i|X=x}$ is trained on the set $\{(x_n, a_n)\}_{n=1}^N$ where $a_n = \mathbb{1}_{(y_n=i)}$. In the COVA case, we build $Q^{\mathrm{COVA}}_{Y|X=x}$ using the set $\{q^{\mathrm{cond}}_{A_i|X=x}\}_{i=0}^{K-2}$ according to (13). For each $i = 0, \ldots, K-2$, the classifier $q^{\mathrm{cond}}_{A_i|X=x}$ is trained on the set $\{(x_n, a_n)\}_{n:y_n \ge i}$ (namely, only on the pairs for which $y_n \ge i$) and $a_n = \mathbb{1}_{(y_n=i)}$.

IV. Leveraging COVA in the Multiclass Case

Let us now present a method that incorporates COVA to reduce the regret of a multiclass classifier. The majority of classifiers in use are parametric. That is, the conditional distribution $Q_{Y|X}$ they output after training is dictated by a vector of parameters, $\theta \in \Theta$. We denote the conditional distribution corresponding to such a classifier by $Q_{Y|X;\theta}$. This distribution induces the conditional binary classifiers

$$q^{\mathrm{cond}}_{A_i|X=x;\theta} = \frac{Q_{Y|X=x;\theta}(i)}{\sum_{j=i}^{K-1} Q_{Y|X=x;\theta}(j)}, \tag{14}$$

which are obviously also fully determined by the parameter vector $\theta \in \Theta$. Thus, each of the induced binary classifiers also belong to a parametric family. Noting that

Scenario	N	Softmax		OVA		COVA		LCOVA	
		Train	Test	Train	Test	Train	Test	Train	Test
A	10^5	-0.005	0.005	0.001	0.009	0.010	0.019	-0.008	0.032
A	10^6	-0.000	$+0.000$	0.004	0.005	0.013	0.014	0.016	0.021
B	10^5	0.703	0.728	0.708	0.732	0.716	0.740	0.663	0.717
B	10^6	0.717	0.718	0.721	0.723	0.730	0.731	0.682	0.690

TABLE I
EXPERIMENTAL RESULTS. THE ENTRIES REPRESENTS REGRET VALUES NATURAL LOGARITHM

$Q_{Y|X;\theta}$ can be written as in (13), replacing $q^{\text{cond}}_{A_i|X=x}$ with $q^{\text{cond}}_{A_i|X=x;\theta}$, Corollary 2 implies that

$$R(P_{Y|X}, Q_{Y|X;\theta}|\ P_X) = \sum_{i=0}^{K-2} \Pr(Y \geq i) r(p^{\text{cond}}_{A_i|X}, q^{\text{cond}}_{A_i|X;\theta}|\ P_{X|Y\geq i}).$$

Noting that all binary classifiers in (14) are determined by the *same* parameter vector $\theta \in \Theta$, an approach for improving the classifier immediately becomes apparent: allow to use a different parameter vector $\theta_i \in \Theta$ for each $q^{\text{cond}}_{A_i|X=x;\theta_i}$. If for each i we choose $\theta_i = \operatorname{argmin}_{\theta\in\Theta} r(p^{\text{cond}}_{A_i|X}, q^{\text{cond}}_{A_i|X;\theta}\ |\ P_{X|Y\geq i})$, we are guaranteed to get a smaller (or identical) regret for the obtained multiclass classifier. We therefore propose to separately train each of the binary classifiers $q^{\text{cond}}_{A_i|X=x;\theta_i}$ such as to minimize the empirical loss over the parameter space Θ, and then merge them into a multiclass classifier using the COVA equation (13). We term this method the *leveraged COVA* (LCOVA). Given enough training samples, the LCOVA method is guaranteed to attain a smaller (or identical) regret than the baseline multiclass classifier $Q_{Y|X;\theta}$. The generalization error, on the other hand, might be greater, as we can now use K different parameter vectors, rather than one.

Let us now demonstrate the LCOVA method for the important special case where the baseline multiclass classifier is logistic regression (Softmax), which corresponds to: $Q_{Y|X=x;\theta}(i) = \exp(\beta_i^T x) / \left[\sum_{j=0}^{K-1} \exp(\beta_j^T x)\right]$. The induced ith conditional binary classifier is: $q^{\text{cond}}_{A_i|X=x;\theta_i} = \exp((\beta_i^i)^T x) / \left[\sum_{j=i}^{K-1} \exp((\beta_j^i)^T x)\right]$ whose parameter set is $\theta_i = \{\beta_j^i\}_{j=i}^{K-1}$ (we do not constrain any β_j^i vector to zero). The sets $\{\beta_j^i\}_{j=i}^{K-1}$ for all $i \in \{0, K-1\}$ can be learned by independently minimizing the following empirical log-loss functions:

$$\hat{L}_i(T, \{\beta_j^i\}_{j=i}^{K-1}) = -\sum_{n:y_n=i} (\beta_i^i)^T x_n - \sum_{n:y_n>i} \log\left[\sum_{j=i+1}^{K-1} e^{(\beta_j^i)^T x_n}\right] + \sum_{n:y_n\geq i} \log\left[\sum_{j=i}^{K-1} e^{(\beta_j^i)^T x_n}\right]. \quad (15)$$

This minimization can be carried by standard optimization tools such as stochastic gradient descent (SGD).

V. EXPERIMENTAL RESULTS

Table I presents experimental results for classifiers designed using the OVA, COVA and LCOVA methods, as well as a standard Softmax classifier, applied on synthetic data with $K = 10$ classes of dimension $d = 100$. In the experiments, $Y \sim \text{Uniform}(\{0, \ldots, K-1\})$, and the conditional distribution of X for the kth class is Gaussian: $[X|Y = i] \sim \mathcal{N}(\mu_i, \Sigma_i)$. We experiment on two scenarios. In both, the class centers $\{\mu_i\}$ are distinct. In Scenario A, $\Sigma_i = \sigma^2 \boldsymbol{I}$ and in Scenario B $\{\Sigma_i\}$ are class dependent and non-diagonal. All binary logistic regression and Softmax classifiers were trained using a standard Python implementation [22]. LCOVA classifiers were trained by minimizing the loss in (15) using SGD. Since the data was synthetically generated, we have access to the ground-truth distribution. The regrets in the table were calculated with respect to the optimal Bayes log-losses, which were approximated by a Monte-Carlo simulation using the true distributions .

It is well known, that given enough samples, Softmax can approach the Bayes log-loss in the additive Gaussian case, which is expressed by its near zero regrets in Scenario A. The rest of the classifiers have small regrets, but as expected, do not outperform Softmax. In Scenario B, the regret of Softmax is strictly positive for any number of samples, and we can see that it is outperformed by LCOVA. The susceptibility to the training data size is also observed in this data, and overfitting is manifested by the differences between the training regrets test counterparts. All classifier show some degree of overfitting at $N = 10^5$, which diminishes at $N = 10^6$. It is in-place to note that despite the fact that LCOVA has $\approx K/2$ more parameters than any of the other three classifiers OVA, COVA and Softmax, its generalization error is still quite reasonable, and it outperforms them in both train and test regrets. In DNNs, one usually has a number of training samples far exceeding the number of classes. Consequently, increasing the number of parameters in the model by a factor of $\approx K/2$ should not have a noticeable effect on the generalization error. Studying the effect of replacing the Softmax classifier in the last layer of DNNs with LCOVA, is therefore a promising direction for future research.

VI. ACKNOWLEDGMENT

The authors would like to thank Meir Feder and Yury Polyanskiy for many fruitful discussions. This work was supported by the ISF under Grant 1791/17.

REFERENCES

[1] N. Merhav and M. Feder, "Universal prediction," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2124–2147, 1998.

[2] N. Cesa-Bianchi and G. Lugosi, *Prediction, learning, and games*. Cambridge University Press, 2006.

[3] T. A. Courtade and T. Weissman, "Multiterminal source coding under logarithmic loss," *IEEE Transactions on Information Theory*, vol. 60, no. 1, pp. 740–761, 2014.

[4] J. Jiao, T. A. Courtade, K. Venkat, and T. Weissman, "Justification of logarithmic loss via the benefit of side information," *IEEE Transactions on Information Theory*, vol. 61, no. 10, pp. 5357–5365, October 2015.

[5] Y. Fogel and M. Feder, "Universal batch learning with log-loss," in *2018 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2018, pp. 21–25.

[6] A. Xu and M. Raginsky, "Minimum excess risk in bayesian learning," *arXiv preprint arXiv:2012.14868*, 2020.

[7] A. Daniely, S. Sabato, and S. S. Shwartz, "Multiclass learning approaches: A theoretical comparison with implications," *arXiv preprint arXiv:1205.6432*, 2012.

[8] R. Rifkin and A. Klautau, "In defense of one-vs-all classification," *The Journal of Machine Learning Research*, vol. 5, pp. 101–141, 2004.

[9] T. Hastie, R. Tibshirani *et al.*, "Classification by pairwise coupling," *Annals of statistics*, vol. 26, no. 2, pp. 451–471, 1998.

[10] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of artificial intelligence research*, vol. 2, pp. 263–286, 1994.

[11] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," *Journal of machine learning research*, vol. 1, no. Dec, pp. 113–141, 2000.

[12] S. Kumar, J. Ghosh, and M. M. Crawford, "Hierarchical fusion of multiple classifiers for hyperspectral data analysis," *Pattern Analysis & Applications*, vol. 5, no. 2, pp. 210–220, 2002.

[13] Y. Chen, M. M. Crawford, and J. Ghosh, "Integrating support vector machines in a hierarchical output space decomposition framework," in *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, vol. 2. IEEE, 2004, pp. 949–952.

[14] V. Vural and J. G. Dy, "A hierarchical method for multi-class support vector machines," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 105.

[15] A. C. Lorena, A. C. De Carvalho, and J. M. Gama, "A review on the combination of binary classifiers in multiclass problems," *Artificial Intelligence Review*, vol. 30, no. 1-4, p. 19, 2008.

[16] P. del Moral, S. Nowaczyk, A. Sant'Anna, and S. Pashami, "Pitfalls of assessing extracted hierarchies for multi-class classification," *arXiv preprint arXiv:2101.11095*, 2021.

[17] A. Bhatt, B. Nazer, O. Ordentlich, and Y. Polyanskiy, "Information-distilling quantizers," *IEEE Trans. Inf. Theory*, accepted, 2021.

[18] O. Ordentlich and I. Tal, "An upgrading algorithm with optimal power law," *arXiv preprint arXiv:2004.00869*, 2020.

[19] A. Beygelzimer, J. Langford, and P. Ravikumar, "Multiclass classification with filter trees," *Preprint, June*, vol. 2, 2007.

[20] A. Ben-Yishai and O. Ordentlich, "Constructing multiclass classifiers using binary classifiers under log-loss," *arXiv preprint arXiv:2102.08184*, 2021.

[21] T. M. Cover and J. Thomas, *Elements of Information Theory*, 2nd ed. New York: Wiley, 2006.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, https://scikit-learn.org/stable/.