

Boosting Classifiers with Noisy Inference

Yongjune Kim, Yuval Cassuto, and Lav R. Varshney

Abstract—We present a principled framework to address resource allocation for realizing boosting algorithms on substrates with communication or computation noise. Boosting classifiers (e.g., AdaBoost) make a final decision via a weighted vote from the outputs of many base classifiers (weak classifiers). Suppose that the base classifiers’ outputs are noisy or communicated over noisy channels; these noisy outputs will degrade the final classification accuracy. We show that this degradation can be effectively reduced by allocating more system resources for more important base classifiers. We formulate resource optimization problems in terms of importance metrics for boosting. Moreover, we show that the optimized noisy boosting classifiers can be more robust than bagging for the noise during inference (test stage). We provide numerical evidence to demonstrate the benefits of our approach.

I. INTRODUCTION

Boosting methods are machine learning algorithms that construct a set of base (weak) classifiers and then classify a new data point by taking a *weighted* vote of their decisions [1]. Boosting can achieve good classification accuracy even if the base classifiers have performance that is only slightly better than random guessing [2], [3]. Adaptive boosting (AdaBoost) is the most widely used form of boosting [2], [4]; it works well for classification problems such as face detection [5] and can be extended to regression problems [6].

Consider the standard supervised classification problem. For the given training set $S = \{(x_1, y_1), \dots, (x_N, y_N)\}$, the objective of learning is to estimate the unknown function $y = f(x)$ based on the given training set. The input vector is given by $x_n = (x_{n,1}, \dots, x_{n,D})$ where D denotes the dimension of the input vectors. The output variables y_n are typically drawn from a discrete set of classes, i.e., $y \in \{1, \dots, K\}$ where K denotes the number of classes. For a binary classification problem, we assume $y \in \{+1, -1\}$.

The final output of AdaBoost is as follows:

$$f(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t f_t(x) \right), \quad (1)$$

where $\alpha = (\alpha_1, \dots, \alpha_T)$ denotes the coefficients for base classifiers. AdaBoost assigns larger coefficients to more accurate (or important) base classifiers [3], [4]. Unlike AdaBoost, the coefficients of bagging classifiers are uniform (i.e., $\alpha_t = \frac{1}{T}$)

Y. Kim was with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA. He is now with Western Digital Research, Milpitas, CA 95035 USA (e-mail: yongjune.kim@wdc.com). Y. Cassuto is with the Viterbi Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel (e-mail: ycassuto@ee.technion.ac.il). L. R. Varshney is with the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Urbana, IL 61801 USA (e-mail: varshney@illinois.edu).

Lav R. Varshney was supported in part by the National Science Foundation under Grant CCF-1717530. Yuval Cassuto was supported in part by the US-Israel Binational Science Foundation and the Israel Science Foundation.

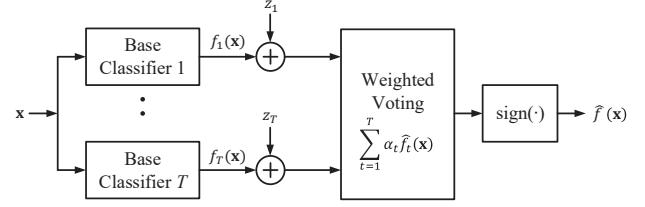


Fig. 1. Noisy boosting classifier.

for all $t \in \{1, \dots, T\}$) and the output of bagging corresponds to majority voting [7].

Suppose that the outputs of base classifiers are corrupted by random noise as shown in Fig. 1. The noise $\mathbf{z} = (z_1, \dots, z_T)$ captures *communication* errors over the channel between the base classifiers and the weighted voter. Alternatively, \mathbf{z} can originate from noise in the *computation* hardware of base classifiers [8]. The corrupted output of the t -th base classifier is denoted by $\hat{f}_t(x) \in \{+1, -1\}$. We assume that the weighted vote is implemented in a noiseless manner.

We observe that noise in the individual base classifiers affects the overall classification accuracy in a way that strongly depends on the coefficient vector α . That is, an erroneous \hat{f}_t with a large coefficient α_t is more likely to corrupt the final classification output than a base classifier with a smaller coefficient. Following this observation, we develop a principled framework to optimize the classification accuracy by allocating reliability resources to base classifiers according to their importance prescribed in the coefficient vector α .

Ideally, the system resources should be allocated to minimize the classification error probability. However, the classification error probability of boosting depends on the data sets and base classifiers (and their training algorithms); the classification error probability is not simply related to the system resources and does not yield tractable optimization procedures. To circumvent this problem, we minimize *proxies* instead of the classification error probability. First, we define three proxies: 1) Markov proxy, 2) Chernoff proxy, and 3) Gaussian proxy. Next, we formulate optimization problems to minimize these proxies for a given resource budget. This kind of indirect approach is effective in many engineering problems, e.g., [9]–[11].

In this paper, we assume that the impact of \mathbf{z} can be controlled by allocating system resources. One example we investigate is that the outputs of base classifiers are corrupted by additive noise over the channels between base classifiers and the weighted voter. Here, the noise level over these channels can be controlled by allocating transmit power. We show the proposed framework can effectively reduce the classification error probability for a given transmit-power budget.

Our approach provides a general framework to allocate a limited resource for boosting classifiers and can also be applied to settings of noisy computations. For example, the quality of computations on noisy hardware can be changed by controlling supply voltage [8], replicating computations [12], [13], and implementing granular bit precisions [10]. Based on the proposed framework, we can optimize these system resources in a principled manner. For the opposite problem, adversarial attackers can exploit the importance metrics to best degrade the classification accuracy. In this attacking scenario, the attackers should allocate more attacking resource to more important base classifiers. We focus on AdaBoost in this paper, but our approach can be applied to any other weighted ensemble method in machine learning.

Our problem of noisy AdaBoost is distinct from AdaBoost in the presence of noisy labels. A well-known model of random classification noise (RCN) assumes that each label y in the training set is flipped independently [4], [5]. Several studies have investigated the behavior of AdaBoost under label noise and proposed more robust training algorithms [16]–[18]. Note that the *data* noise affects all base classifiers during training; hence, it affects the AdaBoost model (i.e., base classifiers and their coefficients) permanently. Our model assumes that the *system* noise during inference (test stage) affects the decisions of base classifiers independently. We optimize the system resource to mitigate the noise impact without altering the trained AdaBoost models.

It is well known that the classification accuracy of AdaBoost tends to degrade more than that of bagging for the RCN model [16], [19]. The reason is that AdaBoost more aggressively fits noisy instances in the training set [15], [16]. Contrarily, we show that AdaBoost can be more robust than bagging in our problem setting where noise flips the base classifiers' outputs during inference (test). This is because the accuracy improvement by the proposed optimization is more effective as the coefficient variability increases, or formally, as the geometric mean of the coefficients decreases.

Our noise model is different from the model in *error-aware* inference [20], [21]. Error-aware inference is a retraining approach to overcome computational errors due to hardware non-idealities. The basic idea is to retrain (i.e., update the trained models) by taking into account noisy computations [8]. However, this retraining approach only works for permanent noise (e.g., stuck-at faults) since random (transient) noise cannot be trained. In this paper, we focus on random errors.

If we regard base classifiers as distributed sensors, then the AdaBoost's weighted voting is similar to the Bayesian fusion rule of *distributed detection* with multiple sensors [22], [23]. In distributed detection, each local sensor's optimal decision rule is the likelihood ratio test (LRT) for conditionally independent sensor observations [23], and usually the observation of each local sensor is a scalar value. On the other hand, AdaBoost can use any learning algorithms that train their models based on high-dimensional data sets. For example, decision trees [5], [16], support vector machines [24], and neural networks [25], [26] have been investigated for base classifiers of AdaBoost.

Our noisy inference problem is distinct from *channel-aware* distributed detection with multiple sensors [27], [28]. The

distinction lies in that our approach controls the channel distributions whereas channel-aware distributed detection attempts to optimize the thresholds of each sensor's LRT for fixed channel distributions. In addition, the approach of channel-aware distributed detection does not work for AdaBoost because 1) the distribution of \mathbf{x} is not known in learning problems, 2) base classifiers are not simple LRTs, and 3) AdaBoost does not require the conditional independence assumption among base classifiers unlike distributed detection with multiple sensors.

The rest of this paper is organized as follows. Section II explains the noisy AdaBoost model. Section III develops three metrics for the importance of base classifiers originating from three optimization problems. Section IV formulates and solves resource-allocation problems based on these importance metrics. Section V provides numerical results and Section VI concludes.

II. NOISY ADABOOST MODEL

A. AdaBoost

AdaBoost trains the base classifiers in sequence to minimize an exponential error function [3], [4]. Each base classifier is trained using a weighted form of the training set in which the data weights $\mathbf{w} = (w_1, \dots, w_N)$ depend on the performance of previous base classifiers. In particular, data points that are misclassified by one of the base classifiers are given greater weight when used to train the next base classifier. Once all base classifiers have been trained, their outputs are combined through weighted voting [3].

Note that the data weights $\mathbf{w} = (w_1, \dots, w_N)$ are distinct from the classifier coefficients $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_T)$. AdaBoost determines both values during training. Once training is done, only the coefficients $\boldsymbol{\alpha}$ are used to classify new data points. The training of AdaBoost is given by Algorithm I.

The final model of AdaBoost is given by (1) where the base classifiers and the coefficients are decided by Algorithm I. The classification error probability of the trained model $f(\cdot)$ is given by

$$P_{e,f} = \Pr(f(\mathbf{x}) \neq y), \quad (6)$$

where y is the true label corresponding to \mathbf{x} .

Remark 1 (Positive Coefficients): If a base classifier is better than random guessing, then $\alpha_t > 0$ for any $t \in \{1, \dots, T\}$ [2].

Remark 2 (Normalized Coefficients): We normalize the coefficients such that $\sum_{t=1}^T \alpha_t = 1$. Note that normalization does not affect the classification output in (1).

Remark 3 (Distinction from Distributed Detection Problem): We note that ε_t in (3) depends on the data weights \mathbf{w} unlike the distributed detection problem in [23].

B. Noisy AdaBoost

Suppose that the base classifiers' outputs may be flipped due to the noise z_t , i.e., $f_t(\mathbf{x}) \neq \hat{f}_t(\mathbf{x})$ where

$$\hat{f}_t(\mathbf{x}) = \text{sign}(f_t(\mathbf{x}) + z_t). \quad (7)$$

The mismatch event of the t -th base classifier is denoted by

$$\delta_t = \mathcal{I}(f_t(\mathbf{x}) \neq \hat{f}_t(\mathbf{x})). \quad (8)$$

Algorithm 1 Training of AdaBoost for binary classification [3]

- 1: Initialize the data weights \mathbf{w} by setting $w_n^{(1)} = \frac{1}{N}$.
- 2: **for** $t = 1 : T$ **do**
- 3: Fit a base classifier $f_t(\mathbf{x})$ to the training set by minimizing the weighted error function

$$J_t = \sum_{n=1}^N w_n^{(t)} \mathcal{I}(f_t(\mathbf{x}_n) \neq y_n) \quad (2)$$

where $\mathcal{I}(f_t(\mathbf{x}_n) \neq y_n)$ denotes the indicator function and equals 1 if $f_t(\mathbf{x}_n) \neq y_n$ and 0 otherwise.

- 4: Evaluate
- 5: Compute the classifier coefficients
- 6: Update the data weights
- 7: **end for**
- 8: **return** the trained base classifiers $\{f_t(\cdot)\}$ and the corresponding coefficients α for $t \in \{1, \dots, T\}$.

Then, we can define the base classifiers' mismatch probabilities as $\mathbf{p} = (p_1, \dots, p_T)$ where

$$p_t \triangleq \Pr(f_t(\mathbf{x}) \neq \hat{f}_t(\mathbf{x})) = \mathbb{E}[\delta_t]. \quad (9)$$

In the sequel, the expectation over the distribution of \mathbf{x} will be replaced by the empirical mean over the given data set.

The final output of noisy AdaBoost is given by

$$\hat{f}(\mathbf{x}) = \text{sign} \left(\sum_{t=1}^T \alpha_t \hat{f}_t(\mathbf{x}) \right). \quad (10)$$

Then, the final mismatch probability (i.e., mismatch probability of the final output) is given by

$$P_m \triangleq \Pr(f(\mathbf{x}) \neq \hat{f}(\mathbf{x})), \quad (11)$$

which captures the negative impact of \mathbf{z} on the final classification accuracy. We can expect that the final mismatch probability P_m depends on the base classifiers' mismatch probabilities \mathbf{p} .

The classification error probability of the noisy AdaBoost is upper bounded by

$$P_e = P_{e,\hat{f}} \leq P_{e,f} + P_m, \quad (12)$$

where $P_{e,f}$ denotes the classification error probability by the noise-free AdaBoost. In [10], [11], the mismatch probability characterizes the impact of quantization noise due to limited bit precision. Note that $P_{e,f}$ solely depends on the AdaBoost algorithm and the dataset, i.e., $P_{e,f}$ is independent of \mathbf{z} . Hence, we focus on P_m to reduce the negative impact of the noise \mathbf{z} .

III. IMPORTANCE METRICS OF BASE CLASSIFIERS

We define three proxies to the mismatch probability: 1) Markov proxy, 2) Chernoff proxy, and 3) Gaussian proxy. These proxies induce different importance metrics of base classifiers. We provide theoretical justification for the proxies and the corresponding metrics.

A. Markov Proxy

Here, we define the *Markov proxy*, which comes from Markov's inequality.

Definition 1 (Markov Proxy): The Markov proxy \hat{p}_M of the noisy AdaBoost is given by

$$\hat{p}_M = \sum_{t=1}^T \alpha_t p_t, \quad (13)$$

which is the nonnegative weighted sum of p_t .

We derive an upper bound on the mismatch probability P_m based on Markov's inequality and show that this upper bound can be lowered by minimizing the Markov proxy \hat{p}_M .

Theorem 1 (Upper Bound by Markov's Inequality): The mismatch probability of \mathbf{x}_n is upper bounded as follows:

$$P_m(\mathbf{x}_n) \leq \frac{2\hat{p}_M}{\gamma_n}, \quad (14)$$

where

$$\gamma_n = \left| \sum_{t=1}^T \alpha_t f_t(\mathbf{x}_n) \right|, \quad (15)$$

which represents the decision margin of \mathbf{x}_n . Then, an upper bound on the mismatch probability P_m is given by

$$P_m \leq \left(\frac{2}{N} \sum_{n=1}^N \frac{1}{\gamma_n} \right) \cdot \hat{p}_M. \quad (16)$$

Proof: The proof is given in Appendix A ■

Higher decision margin γ_n and/or lower Markov proxy \hat{p}_M reduce the upper bound on the mismatch probability. The margin γ_n depends only on the input vector \mathbf{x}_n and the trained AdaBoost classifier model. In contrast, \hat{p}_M depends on \mathbf{z} , whose distribution we can control by resource allocation; hence minimizing \hat{p}_M is pursued in Section IV.

Remark 4: For a given dataset and trained AdaBoost classifier model, the upper bound (14) depends only on the Markov proxy \hat{p}_M . Hence, our objective in Section IV is to minimize \hat{p}_M by controlling $\mathbf{p} = (p_1, \dots, p_T)$.

B. Chernoff Proxy

In this subsection, we define the *Chernoff proxy* via the Chernoff bound.

Definition 2 (Chernoff Proxy): The Chernoff proxy \hat{p}_C of the mismatch probability is given by

$$\hat{p}_C(s) = \sum_{t=1}^T (e^{s\alpha_t} - 1) p_t, \quad (17)$$

where $s > 0$ is a parameter.

Remark 5: Since $\alpha_t > 0$ and $s > 0$, $e^{s\alpha_t} - 1 > 0$.

We derive an upper bound on the mismatch probability P_m from the Chernoff bound and show that this upper bound can be reduced by minimizing the Chernoff proxy \hat{p}_C .

Theorem 2 (Upper Bound by Chernoff Bound): The mismatch probability is upper bounded as follows:

$$P_m \leq \mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right] \cdot \exp(\hat{p}_C(s)), \quad (18)$$

for any $s > 0$. Note that $\mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right]$ can be calculated by

$$\mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right] = \frac{1}{N} \sum_{n=1}^N \exp \left(-s \cdot \frac{\gamma_n}{2} \right). \quad (19)$$

Proof: The proof is given in Appendix B. \blacksquare

This upper bound can be tightened by minimizing the Chernoff proxy \hat{p}_C . Also, similarly to the Markov proxy, a higher decision margin γ_n decreases the upper bound.

In addition, s should be carefully chosen because of a trade-off relation between $\mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right]$ and \hat{p}_C . A smaller s decreases \hat{p}_C while increasing $\mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right]$. Since the optimal s and \mathbf{p} are interdependent, we propose an iterative algorithm to jointly find the optimal s and \mathbf{p} (see Algorithm 2 in Section IV).

C. Gaussian Approximation

As in Definition 1 and Definition 2, we define a third proxy metric that is a nonnegative weighted sum of p_t .

Definition 3 (Gaussian Proxy): The Gaussian proxy \hat{p}_G of the mismatch probability is given by

$$\hat{p}_G = \sum_{t=1}^T \alpha_t^2 p_t. \quad (20)$$

Suppose that $\hat{f}(\mathbf{x}_n) = \text{sign}(\hat{g}(\mathbf{x}_n))$ where $\hat{g}(\mathbf{x}_n)$ is given by

$$\hat{g}(\mathbf{x}_n) = \sum_{t=1}^T \alpha_t \hat{f}_t(\mathbf{x}_n) \quad (21)$$

$$= \sum_{t \in \mathcal{T}_n^+} \alpha_t \hat{f}_t(\mathbf{x}_n) + \sum_{t \in \mathcal{T}_n^-} \alpha_t \hat{f}_t(\mathbf{x}_n) \quad (22)$$

$$= \sum_{t \in \mathcal{T}_n^+} \alpha_t (1 - 2\delta_{t,n}) + \sum_{t \in \mathcal{T}_n^-} \alpha_t (-1 + 2\delta_{t,n}) \quad (23)$$

$$= \pm \gamma_n + v_n \quad (24)$$

where $\mathcal{T}_n^+ = \{t \mid f_t(\mathbf{x}_n) = 1\}$ and $\mathcal{T}_n^- = \{t \mid f_t(\mathbf{x}_n) = -1\}$, respectively. Note that $\delta_{t,n} = \mathcal{I}(f_t(\mathbf{x}_n) \neq \hat{f}_t(\mathbf{x}_n))$.

We set $\pm \gamma_n$ and v_n in (24) as *signal* term and the *noise* term, respectively. The noise term v_n is given by

$$v_n = -2 \left(\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right). \quad (25)$$

TABLE I
COMPARISON OF IMPORTANCE METRICS OF BASE CLASSIFIERS

Proxy	Importance metric β	Remarks
Markov	α	Definition 1
Chernoff	$e^{s\alpha} - 1$	Definition 2
Gaussian	α^2	Definition 3

Theorem 3: The noise term v_n for $n \in \{1, \dots, N\}$ can be modeled as a Gaussian distribution, i.e., $v_n \sim \mathcal{N}(\mu_v, \sigma_v^2)$ by the central limit theorem. Then,

$$\mu_v = -2 \left(\sum_{t \in \mathcal{T}_n^+} \alpha_t p_t - \sum_{t \in \mathcal{T}_n^-} \alpha_t p_t \right), \quad (26)$$

$$\sigma_v^2 = 4 \sum_{t=1}^T \alpha_t^2 p_t (1 - p_t). \quad (27)$$

Proof: The proof is given in Appendix D. \blacksquare

Observe that the variance in (27) is *data independent*, and thus its minimization by resource allocation is an effective way to reduce the classification error probability.

Remark 6: If $p_t \ll 1$ (i.e., $p_t^2 \ll p_t$), then $\sigma_v^2 \approx 4\hat{p}_G$. The advantage of Gaussian proxy is the convexity of the right-hand side of (20) (see Section IV).

Based on the Gaussian approximation, we can derive an estimate of the mismatch probability P_m .

Corollary 4: An estimate of the mismatch probability is given by

$$P_m(\mathbf{x}_n) \approx Q \left(\frac{\gamma_n - \mu_v}{2\sqrt{\hat{p}_G}} \right), \quad (28)$$

$$\text{where } Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp \left(-\frac{u^2}{2} \right) du.$$

The minimum μ_v and the minimum \hat{p}_G are desired to reduce the estimate of mismatch probability. However, μ_v depends on the input vector \mathbf{x}_n , hence, we cannot easily minimize μ_v . In contrast, the Gaussian proxy depends only on the trained α and the base classifiers' mismatch probability \mathbf{p} . Hence, we can minimize the Gaussian proxy to reduce the mismatch probability.

Note that each of the three proxies can be described by

$$\sum_{t=1}^T \beta_t p_t, \quad (29)$$

where β_t denotes the *importance metric* of the t -th base classifier. Table I lists the importance metrics for the three proxies.

Fig. 2 plots the dependence of each importance metric on α . It illustrates how resource allocation based on β would give preference to a larger α . Note that the importance metrics of Markov proxy and Gaussian proxy correspond to ℓ_1 -norm penalty and ℓ_2 -norm penalty, respectively. Hence, the Gaussian proxy puts very small weight on less important base classifiers, but strong weight on more important base classifiers.

Remark 7: The proposed importance metrics are positive (i.e., $\beta_t > 0$) because of $\alpha_t > 0$ (Remark 1) and $e^{s\alpha_t} - 1 > 0$ (Remark 5).

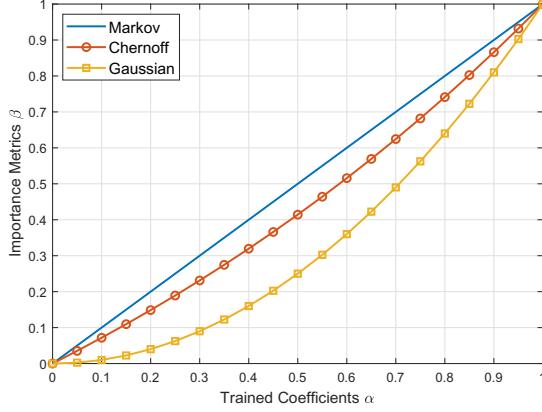


Fig. 2. Comparison of importance metrics β in Table II (with $s = \log 2$ for the Chernoff proxy).

IV. RESOURCE ALLOCATION FOR NOISY ADABoost

A. Formulation of Optimization Problems

We investigate optimization approaches to determine the optimal $\mathbf{p} = (p_1, \dots, p_T)$ for a given resource constraint. By optimizing the proposed proxies, we attempt to reduce the mismatch probability, i.e., reduce the noise impact on classification accuracy.

An important assumption is that the mismatch probabilities of base classifiers can be controlled by allocating the resources. Suppose that the mismatch probability of the t -th base classifier p_t can be described by resource r_t , i.e., $p_t = p(r_t)$. Then, we can formulate the following optimization problem for a given resource budget \mathcal{C} :

$$\begin{aligned} & \underset{\mathbf{r}}{\text{minimize}} \quad \sum_{t=1}^T \beta_t p(r_t) \\ & \text{subject to} \quad \sum_{t=1}^T c(r_t) \leq \mathcal{C} \end{aligned} \quad (30)$$

where the objective function depends on the importance metric $\beta = (\beta_1, \dots, \beta_T)$. Also, $c(r_t)$ denotes the cost of the allocated resource to the t -th base classifier.

Remark 8: If $p(r_t)$ and $c(r_t)$ are convex, then the optimization problem (30) is also convex because β_t is positive for all t in any of the three proxies (Remark 7). In such cases, for the Markov proxy and the Gaussian proxy, we can obtain the optimal resource allocation by solving (30) directly using convex programming.

For the Chernoff proxy (17), due to the free parameter s , we propose an iterative algorithm to jointly find the optimal s and \mathbf{p} (see Algorithm 2).

Algorithm 2 attempts to minimize the upper bound of (18) by alternating between optimizations of (31) and (32). Step 1 finds $\mathbf{p}^{(i+1)}$ minimizing $\hat{p}_C(s^{(i)})$ for a given $s^{(i)}$, which is a convex problem for any $s^{(i)} > 0$. It is because \mathbf{p} affects only $\hat{p}_C(s)$ among the upper bound of (18), which corresponds to (30). Step 2 finds in closed form $s^{(i+1)}$ minimizing the upper bound of (18) for a given $\mathbf{p}^{(i+1)}$. Step 3 introduces a small positive $\epsilon > 0$ to satisfy the condition of $s > 0$.

Algorithm 2 Iterative algorithm to minimize the upper bound in Theorem 2

- 1: Choose an arbitrary starting point $s^{(0)}$ in $(0, \infty)$ and set $i = 0$.
- 2: **repeat**
- 3: Step 1. Solve the following optimization problem:
- $$\mathbf{p}^{(i+1)} = \arg \min_{\mathbf{p}} \hat{p}_C(s^{(i)}). \quad (31)$$
- 4: Step 2. Find $s^{(i+1)}$ satisfying
- $$\sum_{n=1}^N \left\{ \left(\sum_{t=1}^T p_t^{(i+1)} \alpha_t e^{s \alpha_t} - \frac{\gamma_n}{2} \right) \cdot e^{-s \cdot \frac{\gamma_n}{2}} \right\} = 0. \quad (32)$$
- 5: Step 3. $s^{(i+1)} = \max \{s^{(i+1)}, \epsilon\}$.
- 6: Step 4. $i \leftarrow i + 1$.
- 7: **until** stopping criterion is satisfied.

We show that the upper bound is a convex function of s and the solution of (32) is optimal.

Theorem 5: For given $\mathbf{p} = (p_1, \dots, p_T)$ and $\gamma = (\gamma_1, \dots, \gamma_N)$, the upper bound of (18) is a function of s as follows:

$$h(s) = \frac{1}{N} \sum_{n=1}^N e^{-s \cdot \frac{\gamma_n}{2}} \cdot \exp \left(\sum_{t=1}^T (e^{s \alpha_t} - 1) p_t \right). \quad (33)$$

This upper bound function $h(s)$ is convex; hence s satisfying $h'(s) = 0$ (i.e., (32)) is optimal.

Proof: The proof is given in Appendix C. ■

Corollary 6: If $\gamma_n = \gamma_0$ for all $n \in \{1, \dots, N\}$ and $\alpha_t = \frac{1}{T}$ for all $t \in \{1, \dots, T\}$, then the optimal s is given by

$$s^* = \max \left\{ T \cdot \log \frac{\gamma_0}{2 \cdot \tilde{p}}, \epsilon \right\} \quad (34)$$

where $\tilde{p} = \frac{\sum_{t=1}^T p_t}{T}$.

Proof: If $\gamma_n = \gamma_0$ for any n and $\alpha_t = \frac{1}{T}$ for any t , then (33) is given by

$$h(s) = \exp \left(-\frac{\gamma_0}{2} s + \sum_{t=1}^T (e^{\frac{s}{T}} - 1) p_t \right). \quad (35)$$

Then, the minimization of $h(s)$ is equivalent to minimizing $q(s) = -\frac{\gamma_0}{2} s + \sum_{t=1}^T (e^{\frac{s}{T}} - 1) p_t$. It is clear that $s^* = T \cdot \log \frac{\gamma_0}{2 \cdot \tilde{p}}$ satisfies $q'(s) = 0$. If $s^* \leq 0$, then it is replaced by ϵ . ■

Note that a larger noise margin γ_0 increases the optimal s whereas a larger \tilde{p} reduces the optimal s .

B. Example: Communication Power Allocation

Suppose that the $f_t(\mathbf{x}) \in \{+1, -1\}$ is transmitted by a symbol from $\{r_t, -r_t\}$, which is corrupted by the noise z_t as shown in Fig. 1. In many applications, the additive noise can be modeled as Gaussian distribution, i.e., $z_t \sim \mathcal{N}(0, \sigma_t^2)$. Then,

$$p(r_t) = Q \left(\sqrt{\text{SNR}_t} \right) = Q \left(\frac{r_t}{\sigma_t} \right), \quad (36)$$

where the signal-to-noise ratio (SNR) is $\text{SNR}_t = \frac{r_t^2}{\sigma_t^2}$. Hence, p_t can be controlled by allocating transmit power $c(r_t) = r_t^2$. Then, the optimization problem (30) will be

$$\begin{aligned} \text{minimize}_{\mathbf{r}} \quad & \sum_{t=1}^T \beta_t Q\left(\frac{r_t}{\sigma_t}\right) \\ \text{subject to} \quad & \sum_{t=1}^T r_t^2 \leq \mathcal{C} \\ & r_t \geq 0 \quad t = 1, \dots, T \end{aligned} \quad (37)$$

where \mathcal{C} represents the total power budget.

Remark 9: The power allocation problem (37) is a convex optimization problem since $p(r_t) = Q(r_t)$ is convex for $r_t \geq 0$. Note that $\frac{d^2 Q(x)}{dx^2} = \frac{x}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \geq 0$.

Theorem 7: The optimal solution \mathbf{r}^* of (37) is

$$\text{SNR}_t^* = \left(\frac{r_t^*}{\sigma_t}\right)^2 = W\left(\frac{\beta_t^2}{8\pi\sigma_t^4\nu^2}\right), \quad (38)$$

where ν is a positive dual variable of the Karush-Kuhn-Tucker (KKT) conditions. Also, $W(\cdot)$ denotes the *Lambert W function* (i.e., the inverse function of $f(x) = xe^x$) [29].

Proof: The proof is given in Appendix E. \blacksquare

Since $W(x)$ is an increasing function for $x \geq 0$, (38) shows that as desired we allocate higher SNR for higher β_t . For a classifier with $\beta_t \simeq 0$, the corresponding SNR is $\text{SNR}_t^* \simeq 0$ because $W(0) = 0$.

Corollary 8: If $\sigma_t = \sigma$ for all $t \in \{1, \dots, T\}$, then the optimized proxy of (37) can be approximated as:

$$\sum_{t=1}^T \beta_t Q\left(\frac{r_t^*}{\sigma_t}\right) \approx \frac{T}{2} \exp\left(-\frac{\mathcal{C}}{2T\sigma^2}\right) \left(\prod_{t=1}^T \beta_t\right)^{\frac{1}{T}} \quad (39)$$

where $\left(\prod_{t=1}^T \beta_t\right)^{\frac{1}{T}}$ is the geometric mean of β .

Proof: The proof is given in Appendix E. \blacksquare

We observe that a smaller geometric mean of β implies a lower proxy value. Note that higher SNR $\frac{\mathcal{C}}{\sigma^2}$ decreases the proxy value.

Remark 10: The geometric mean of β is maximized for the uniform $\alpha = (\frac{1}{T}, \dots, \frac{1}{T})$. Thus the non-uniform coefficients of AdaBoost's classifiers contribute to lower classification error probability. This suggests an advantage of AdaBoost over bagging that assigns the same coefficients to all classifiers (i.e., $\alpha = (\frac{1}{T}, \dots, \frac{1}{T})$). This is a noteworthy fact because AdaBoost is known to be less robust than bagging in the problem of noisy data labels [15], [16].

V. NUMERICAL RESULTS

We validate the tools and analytic results of Sections III and IV with the UCI breast cancer dataset [30]. We compare mismatch probabilities and classification error probabilities of uniform resource allocation and optimized resource allocations for Markov proxy, Chernoff proxy, and Gaussian proxy. The noise-free AdaBoost was trained by Algorithm 1 with decision stumps as base classifiers. Based on the training output α , we compute β as shown in Table I and solve the corresponding optimization problems by (37).

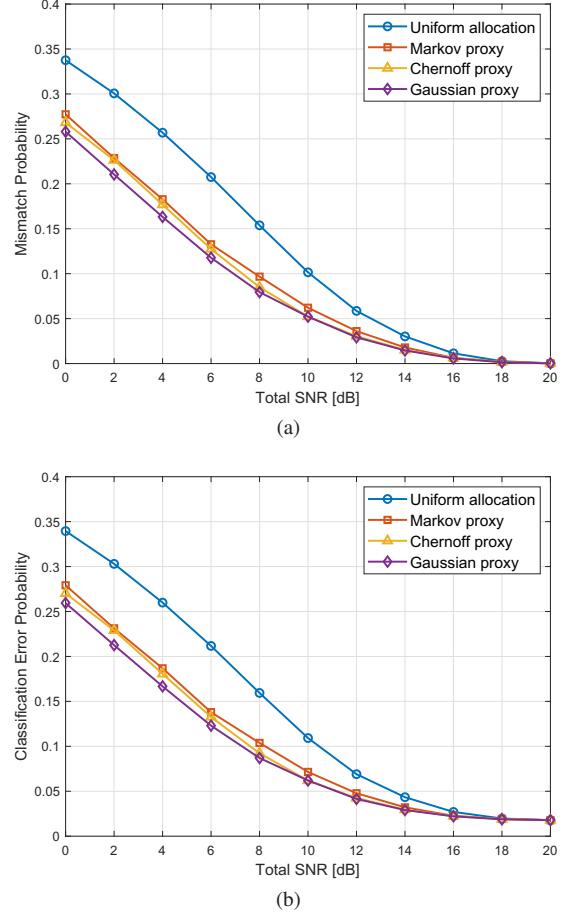


Fig. 3. Evaluation of optimized SNRs ($T = 10$): (a) Mismatch probability and (b) classification error probability.

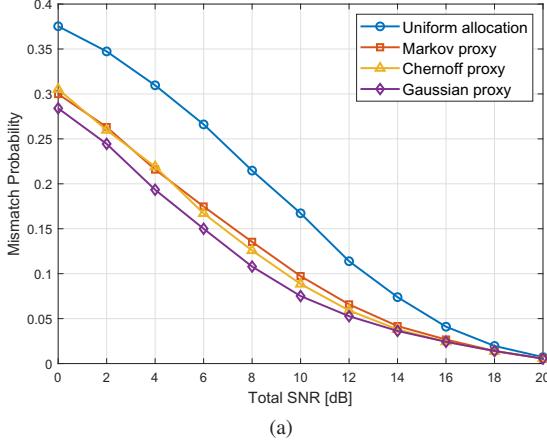
Fig. 3 and Fig. 4 evaluate the mismatch probabilities and the classification error probabilities of the test set for $T = 10$ and $T = 20$, respectively. We observe that nonuniform communication power allocations can lower the mismatch probability as well as the classification error probability. Among the three nonuniform power allocations (Markov proxy, Chernoff proxy, and Gaussian proxy), the power allocation based on Gaussian proxy achieves the best performance. We emphasize that Fig. 3 and Fig. 4 plot the actual mismatch and classification error probabilities over the test set optimized with different proxies, and *not* the values of the proxies themselves. Note that the horizontal axis corresponds to the total SNR budget $\frac{\mathcal{C}}{\sigma^2}$.

For $T = 10$, the Gaussian proxy allocation improves the SNR over the uniform allocation by 3.5 dB at $P_e = 0.1$. For $T = 20$, the SNR gain is 4.2 dB at $P_e = 0.1$. For higher SNR, the mismatch probabilities converge to zero.

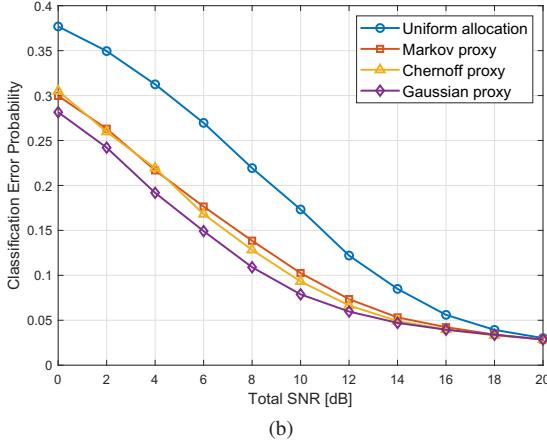
The optimization results by Chernoff proxy are close to the results by Markov proxy in the low-SNR regime. As the SNR increases, the optimization results by Chernoff proxy get close to the results by Gaussian proxy. It can be explained by the Taylor approximation of $\beta = e^{s\alpha} - 1$ in the Chernoff proxy (in Table II) as follows:

$$e^{s\alpha} - 1 \simeq s\alpha + (s\alpha)^2. \quad (40)$$

We observe that the optimal s^* in Algorithm 2 is small for low



(a)



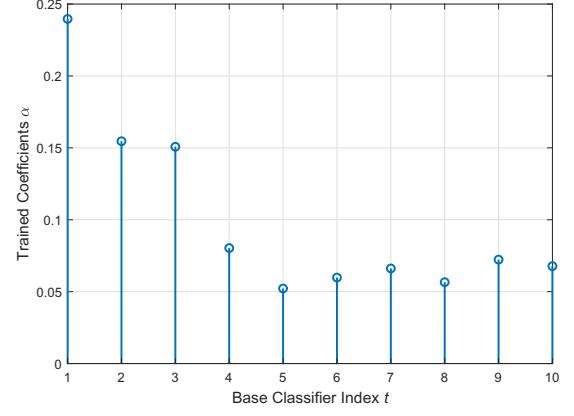
(b)

Fig. 4. Evaluation of optimized SNRs ($T = 20$): (a) Mismatch probability and (b) classification error probability.

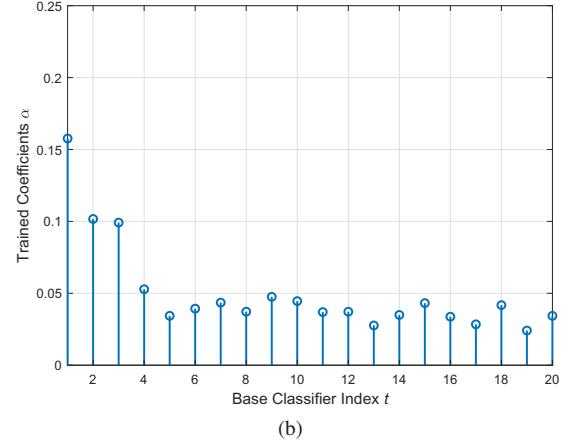
SNR and increases for higher SNR. Hence, β is approximated to $s\alpha$ in the low SNR region because of $s\alpha \ll 1$. As the SNR increases, $(s\alpha)^2$ is a better approximation to β . Then, the corresponding coefficients are $\beta = s^2(\alpha_1^2, \dots, \alpha_T^2)$, which is equivalent to the Gaussian proxy optimization.

Fig. 5 shows the trained coefficients α . We observe that the values of α are nonuniform. Hence, we can improve the mismatch probability and the classification error probability by allocating the optimized-nonuniform transmit power. We expect that the trained AdaBoost models with lower geometric means of β are more robust to the noise as discussed in Corollary 8.

Fig. 6 shows the mismatch probabilities and their estimates for the uniform power allocation and the optimized power allocation using the Gaussian proxy, respectively. The estimates of mismatch probability are calculated by (28). We observe that the estimates match the mismatch probabilities well, which justifies using the Gaussian approximation in Section III-C. On the other hand, the other proxies (Markov and Chernoff) have the advantage that their values are proven upper bounds on the mismatch probability (while the Gaussian proxy in general is not a bound).



(a)



(b)

Fig. 5. Trained coefficients α : (a) $T = 10$ and (b) $T = 20$.

VI. CONCLUSION

In this paper, we propose a principled approach to optimize resource allocation for boosting classifiers. We defined three proxies and the corresponding importance metrics for base classifiers based on Markov inequality, Chernoff bound, and Gaussian approximation. By exploiting the positivity of the importance metrics, we formulated convex resource-allocation problems to minimize the impact of noise. We showed that the proposed approach can effectively improve the classification accuracy for the additive Gaussian noise model. Also, we found that the non-uniform coefficients in boosting offer an advantage over uniform ones (e.g., in bagging) for this noise model. Future work includes different resource/noise models and extension to multi-class noisy inference of boosting classifiers. We believe that the proposed approach can be applied to other settings such as channel-aware distributed detection.

APPENDIX A PROOF OF THEOREM I

The mismatch probability $P_m(\mathbf{x}_n)$ is given by

$$\begin{aligned}
 P_m(\mathbf{x}_n) &= \Pr(f(\mathbf{x}_n) > 0) \Pr(\hat{f}(\mathbf{x}_n) < 0 \mid f(\mathbf{x}_n) > 0) \\
 &\quad + \Pr(f(\mathbf{x}_n) < 0) \Pr(\hat{f}(\mathbf{x}_n) > 0 \mid f(\mathbf{x}_n) < 0). \quad (41)
 \end{aligned}$$

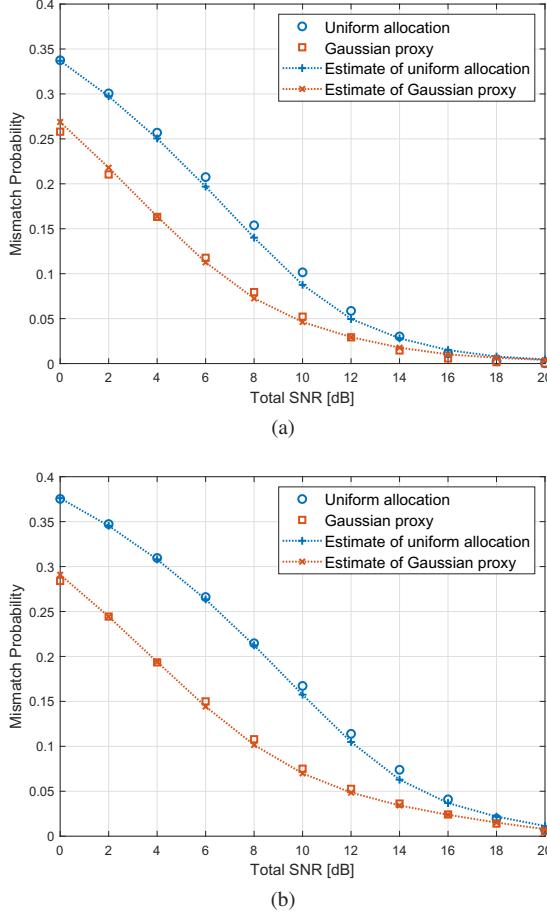


Fig. 6. Mismatch probabilities and their estimates by (28): (a) $T = 10$ and (b) $T = 20$.

Suppose that $g(\mathbf{x}_n) = \sum_{t=1}^T \alpha_t f_t(\mathbf{x}_n) > 0$. By (15),

$$\sum_{t \in \mathcal{T}_n^+} \alpha_t - \sum_{t \in \mathcal{T}_n^-} \alpha_t = \gamma_n. \quad (42)$$

where $\mathcal{T}_n^+ = \{t \mid f_t(\mathbf{x}_n) = 1\}$ and $\mathcal{T}_n^- = \{t \mid f_t(\mathbf{x}_n) = -1\}$, respectively. If $f(\mathbf{x}_n) < 0$, then

$$\sum_{t \in \mathcal{T}_n^+} \alpha_t - \sum_{t \in \mathcal{T}_n^-} \alpha_t = -\gamma_n. \quad (43)$$

By (42) and (23), $\hat{f}(\mathbf{x}_n) < 0$ for given $f(\mathbf{x}_n) = \gamma_n > 0$ is equivalent to

$$\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} > \frac{\gamma_n}{2}. \quad (44)$$

Similarly, $\hat{f}(\mathbf{x}_n) > 0$ for given $f(\mathbf{x}_n) = -\gamma_n < 0$ is equivalent to

$$\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} < -\frac{\gamma_n}{2}. \quad (45)$$

Hence, the mismatch probability is upper bounded by

$$P_m(\mathbf{x}_n) \leq \max \{ \Pr(f(\mathbf{x}_n) > 0), \Pr(f(\mathbf{x}_n) < 0) \} \\ \times \Pr \left(\left| \sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right| > \frac{\gamma_n}{2} \right) \quad (46)$$

$$\leq 2 \cdot \frac{\mathbb{E} \left[\left| \sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right| \right]}{\gamma_n} \quad (47)$$

$$\leq 2 \cdot \frac{\sum_{t=1}^T \alpha_t p_t}{\gamma_n}. \quad (48)$$

where (47) follows from the Markov's inequality and (48) follows from $\mathbb{E}[\delta_{t,n}] = p_t$.

If $\Pr(f(\mathbf{x}_n) > 0) = \Pr(f(\mathbf{x}_n) < 0) = \frac{1}{2}$, then (48) will be

$$P_m(\mathbf{x}_n) \leq \frac{\sum_{t=1}^T \alpha_t p_t}{\gamma_n}. \quad (49)$$

APPENDIX B PROOF OF THEOREM 2

Suppose that $f(\mathbf{x}_n) > 0$. Then, the mismatch probability $p(\mathbf{x}_n)$ is given by

$$P_m(\mathbf{x}_n) = \Pr \left(\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} > \frac{\gamma_n}{2} \right). \quad (50)$$

For any $s > 0$,

$$P_m(\mathbf{x}_n) \\ = \Pr \left(e^{s \left(\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right)} > e^{s \cdot \frac{\gamma_n}{2}} \right) \quad (51)$$

$$\leq \frac{\mathbb{E} \left[\exp \left(s \left(\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right) \right) \right]}{e^{s \cdot \frac{\gamma_n}{2}}} \quad (52)$$

where (52) comes from the Markov's inequality.

Note that $\delta_{t,n}$ for $t \in \{1, \dots, T\}$ are independent because of independent z_t . Hence,

$$\mathbb{E} \left[\exp \left(s \left\{ \sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right\} \right) \right] \quad (53)$$

$$= \prod_{t \in \mathcal{T}_n^+} \mathbb{E} [\exp(s \alpha_t \delta_{t,n})] \cdot \prod_{t \in \mathcal{T}_n^-} \mathbb{E} [\exp(-s \alpha_t \delta_{t,n})]. \quad (54)$$

In addition,

$$\mathbb{E} [\exp(s \alpha_t \delta_{t,n})] = 1 + p_t \{\exp(s \alpha_t) - 1\} \quad (55)$$

$$\leq \exp(p_t (e^{s \alpha_t} - 1)) \quad (56)$$

where (56) follows from $1 + u \leq \exp(u)$ and $u = p_t (e^{s \alpha_t} - 1)$. Similarly, $\mathbb{E} [\exp(-s \alpha_t \delta_{t,n})] \leq \exp(-p_t (1 - e^{-s \alpha_t}))$.

By (52), (54), and (56),

$$\begin{aligned} P_m(\mathbf{x}_n) &\leq \exp \left(-s \cdot \frac{\gamma_n}{2} + \sum_{t \in \mathcal{T}_n^+} p_t (e^{s\alpha_t} - 1) \right. \\ &\quad \left. - \sum_{t \in \mathcal{T}_n^-} p_t (1 - e^{-s\alpha_t}) \right) \end{aligned} \quad (57)$$

$$\leq \exp \left(-s \cdot \frac{\gamma_n}{2} + \sum_{t=1}^T p_t (e^{s\alpha_t} - 1) \right). \quad (58)$$

Similarly, we can obtain the same upper bound for a case of $\sum_{t=1}^T \alpha_t f_t(\mathbf{x}_n) = -\gamma_n < 0$.

Hence, the upper bound on the mismatch probability is given by

$$\begin{aligned} P_m &= \mathbb{E}[P_m(\mathbf{x}_n)] \\ &\leq \mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right] \cdot \exp \left(\sum_{t=1}^T p_t (e^{s\alpha_t} - 1) \right) \end{aligned} \quad (59)$$

$$= \mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right] \cdot \exp(\hat{p}_C) \quad (60)$$

where a larger s reduces $\mathbb{E} \left[\exp \left(-s \cdot \frac{\gamma_n}{2} \right) \right]$ while it increases $\exp(\hat{p}_C)$.

APPENDIX C PROOF OF THEOREM 5

For given $\mathbf{p} = (p_1, \dots, p_T)$ and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_N)$, the upper bound on P_m is given by

$$\begin{aligned} h(s) &= \left(\frac{1}{N} \sum_{n=1}^N e^{-s \cdot \frac{\gamma_n}{2}} \right) \cdot \exp \left(\sum_{t=1}^T (e^{s\alpha_t} - 1) p_t \right) \\ &= \frac{1}{N} \cdot h_1(s) \cdot h_2(s) \end{aligned} \quad (61)$$

where

$$h_1(s) = \sum_{n=1}^N e^{-s \cdot \frac{\gamma_n}{2}}, \quad (62)$$

$$h_2(s) = \exp \left(\sum_{t=1}^T (e^{s\alpha_t} - 1) p_t \right). \quad (63)$$

We show that $h''(s) \geq 0$ to check the convexity of $h(s)$. First, we obtain

$$h'(s) = h'_1(s)h_2(s) + h_1(s)h'_2(s) \quad (64)$$

$$= \sum_{n=1}^N \left\{ \left(\sum_{t=1}^T p_t \alpha_t e^{s\alpha_t} - \frac{\gamma_n}{2} \right) e^{-s \cdot \frac{\gamma_n}{2}} \right\} \cdot h_2(s) \quad (65)$$

where $h_2(s) > 0$ for $s > 0$ and $\alpha_t > 0$. Hence, $h'(s) = 0$ can be achieved by s satisfying

$$\sum_{n=1}^N \left\{ \left(\sum_{t=1}^T p_t \alpha_t e^{s\alpha_t} - \frac{\gamma_n}{2} \right) e^{-s \cdot \frac{\gamma_n}{2}} \right\} = 0 \quad (66)$$

which is equivalent to (32) in Algorithm 2.

The second derivative of $h(s)$ is given by

$$h''(s) = h_2(s) \cdot \sum_{n=1}^N e^{-s \cdot \frac{\gamma_n}{2}} \cdot \frac{h_3(s)}{4} \quad (67)$$

where $h_3(s)$ is

$$\begin{aligned} h_3(s) &= \gamma_n^2 - 4\gamma_n \sum_{t=1}^T p_t \alpha_t e^{s\alpha_t} \\ &\quad + 4 \left\{ \sum_{t=1}^T p_t \alpha_t^2 e^{s\alpha_t} + \left(\sum_{t=1}^T p_t \alpha_t e^{s\alpha_t} \right)^2 \right\} \end{aligned} \quad (68)$$

$$= \left(\gamma_n - 2\gamma_n \sum_{t=1}^T p_t \alpha_t e^{s\alpha_t} \right)^2 + 4 \sum_{t=1}^T p_t \alpha_t^2 e^{s\alpha_t}. \quad (69)$$

Note that $h_2(s) > 0$ and $h_3(s) > 0$ for $s > 0$, $\alpha_t \geq 0$ and $p_t \geq 0$ for any t . Hence, $h''(s) > 0$ and the optimal s should satisfy (66).

APPENDIX D PROOF OF THEOREM 3

By (24), we showed that $\hat{f}(\mathbf{x}_n) = \pm \gamma_n + v_n$. The classification noise term v_n is given by

$$v_n = -2 \left(\sum_{t \in \mathcal{T}_n^+} \alpha_t \delta_{t,n} - \sum_{t \in \mathcal{T}_n^-} \alpha_t \delta_{t,n} \right). \quad (70)$$

Since we assume that z_t s of (7) are independent, $\delta_{t,n}$ s are also independent for a given \mathbf{x}_n . By the central limit theorem, $\sum_{t \in \mathcal{T}_n^+} \alpha_t p_t$ and $\sum_{t \in \mathcal{T}_n^-} \alpha_t p_t$ can be approximated as Gaussian distributions, respectively. Hence, v_n can be modeled as Gaussian distribution as well.

The mean of v_n is readily derived by using $\mathbb{E}[\delta_{t,n}] = p_t$. The variance of v_n is given by

$$\sigma_v^2 = 4 \sum_{t=1}^T \alpha_t^2 \cdot \text{Var}[\delta_{t,n}] \quad (71)$$

$$= 4 \sum_{t=1}^T \alpha_t^2 \cdot p_t(1 - p_t) \quad (72)$$

where $\text{Var}[\delta_{t,n}] = \mathbb{E}[\delta_{t,n}^2] - \mu_v^2 = p_t(1 - p_t)$.

APPENDIX E PROOF OF THEOREM 7

The Lagrangian $L(\mathbf{r}, \nu, \boldsymbol{\lambda})$ of (37) is given by

$$L(\mathbf{r}, \nu, \boldsymbol{\lambda})$$

$$= \sum_{t=1}^T \beta_t Q \left(\frac{r_t}{\sigma_t} \right) + \nu \left(\sum_{t=1}^T r_t^2 - \mathcal{C} \right) - \sum_{t=1}^T \lambda_t r_t. \quad (73)$$

The corresponding KKT conditions are as follows:

$$\sum_{t=1}^T r_t^2 \leq \mathcal{C}, \quad \nu \geq 0, \quad \nu \cdot \left(\sum_{t=1}^T r_t^2 - \mathcal{C} \right) = 0, \quad (74)$$

$$r_t \geq 0, \quad \lambda_t \geq 0, \quad \lambda_t r_t = 0 \quad (75)$$

$$\frac{\partial L}{\partial r_t} = -\frac{\beta_t}{\sqrt{2\pi\sigma_t}} \exp \left(-\frac{r_t^2}{2\sigma_t^2} \right) + 2\nu r_t - \lambda_t = 0 \quad (76)$$

for $t \in \{1, \dots, T\}$.

From (76), λ_t is

$$\lambda_t = 2\nu r_t - \frac{\beta_t}{\sqrt{2\pi\sigma_t}} \exp \left(-\frac{r_t^2}{2\sigma_t^2} \right). \quad (77)$$

If $\nu = 0$. Then, $\lambda_t < 0$, which violates (75) because of $\beta_t > 0$. Hence, we claim that $\nu \neq 0$, which results in $\sum_{t=1}^T r_t^2 = \mathcal{C}$.

If $r_t = 0$, then $\lambda_t = -\frac{\beta_t}{\sqrt{2\pi}\sigma_t}$, which violates (75). Hence, we claim that $r_t > 0$ and $\lambda_t = 0$, i.e.,

$$r_t = \frac{\beta_t}{2\sqrt{2\pi}\sigma_t\nu} \exp\left(-\frac{r_t^2}{2\sigma_t^2}\right), \quad (78)$$

which is equivalent to

$$\frac{r_t^2}{\sigma_t^2} \exp\left(\frac{r_t^2}{\sigma_t^2}\right) = \frac{\beta_t^2}{8\pi\sigma_t^4\nu^2}. \quad (79)$$

By setting $x = \frac{r_t^2}{\sigma_t^2}$, we obtain $x \exp(x) = \frac{\beta_t^2}{8\pi\sigma_t^4\nu^2}$. Hence, $x = \text{SNR}_t = W\left(\frac{\beta_t^2}{8\pi\sigma_t^4\nu^2}\right)$.

APPENDIX F PROOF OF THEOREM 8

By replacing $Q(x)$ with its Chernoff bound $\frac{1}{2} \exp\left(-\frac{x^2}{2}\right)$, the optimization problem (37) will be modified to

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \sum_{t=1}^T \frac{\beta_t}{2} \exp\left(-\frac{x_t^2}{2}\right) \\ & \text{subject to} \quad \sum_{t=1}^T x_t^2 \leq \frac{\mathcal{C}}{\sigma^2} \\ & \quad x_t \geq 0 \quad t = 1, \dots, T \end{aligned} \quad (80)$$

where $x_t = \frac{r_t}{\sigma_t}$. The corresponding Lagrangian is given by

$$\begin{aligned} L(\mathbf{x}, \nu, \boldsymbol{\lambda}) = & \sum_{t=1}^T \frac{\beta_t}{2} \exp\left(-\frac{x_t^2}{2}\right) \\ & + \nu \left(\sum_{t=1}^T x_t^2 - \mathcal{C}' \right) - \sum_{t=1}^T \lambda_t x_t. \end{aligned} \quad (81)$$

From the KKT conditions, we can obtain the following conditions of the optimal x^* (as in Appendix E):

$$\sum_{t=1}^T (x_t^*)^2 = \frac{\mathcal{C}}{\sigma^2} \quad (82)$$

$$\frac{\beta_t}{2} \exp\left(-\frac{(x_t^*)^2}{2}\right) = 2\nu. \quad (83)$$

By (80) and (83), the optimized proxy is given by

$$\sum_{t=1}^T \frac{\beta_t}{2} \exp\left(-\frac{(x_t^*)^2}{2}\right) = 2T\nu. \quad (84)$$

Note that (83) is equivalent to

$$(x_t^*)^2 = \text{SNR}_t^* = 2 \log \frac{\beta_t}{4\nu}. \quad (85)$$

By (82) and (85),

$$\log \prod_{t=1}^T \frac{\beta_t}{4\nu} = \frac{\mathcal{C}}{2\sigma^2}, \quad (86)$$

which leads to $\nu = \frac{1}{4} \exp\left(-\frac{\mathcal{C}}{2T\sigma^2}\right) \left(\prod_{t=1}^T \beta_t\right)^{\frac{1}{T}}$. Then, the optimized proxy is given by

$$2T\nu = \frac{T}{2} \exp\left(-\frac{\mathcal{C}}{2T\sigma^2}\right) \left(\prod_{t=1}^T \beta_t\right)^{\frac{1}{T}}. \quad (87)$$

REFERENCES

- [1] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Int. Workshop Multiple Classifier Syst.*, Dec. 2000, pp. 1–15.
- [2] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Dec. 1997.
- [3] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
- [4] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Jul. 1996, pp. 148–156.
- [5] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition (CVPR)*, Dec. 2001, pp. I–511–I–518.
- [6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001.
- [7] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.
- [8] N. R. Shanbhag, N. Verma, Y. Kim, A. D. Patil, and L. R. Varshney, "Shannon-inspired statistical computing for the nanoscale era," *Proc. IEEE*, vol. 107, no. 1, pp. 90–107, Jan. 2019.
- [9] H. V. Poor and J. B. Thomas, "Applications of Ali-Silvey distance measures in the design of generalized quantizers for binary decision systems," *IEEE Trans. Commun.*, vol. 25, no. 9, pp. 893–900, Sep. 1977.
- [10] C. Sakr, Y. Kim, and N. Shanbhag, "Analytical guarantees on numerical precision of deep neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 3007–3016.
- [11] C. Sakr, A. Patil, S. Zhang, Y. Kim, and N. Shanbhag, "Minimum precision requirements for the SVM-SGD learning algorithm," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Mar. 2017, pp. 1138–1142.
- [12] J. von Neumann, "Probabilistic logics and the synthesis of reliable organisms from unreliable components," *Automata Studies*, vol. 34, pp. 43–98, 1956.
- [13] M. A. Donmez, M. Raginsky, A. C. Singer, and L. R. Varshney, "Cost-performance tradeoffs in unreliable computation architectures," in *Proc. Asilomar Conf. Signals, Syst. Comput.*, Nov. 2016, pp. 215–219.
- [14] D. Angluin and P. Laird, "Learning from noisy examples," *Mach. Learn.*, vol. 2, no. 4, pp. 343–370, Apr. 1988.
- [15] B. Frenay and M. Verleysen, "Classification in the presence of label noise: A survey," *IEEE Trans. Neural Netw.*, vol. 25, no. 5, pp. 845–869, May 2014.
- [16] T. G. Dietterich, "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization," *Mach. Learn.*, vol. 40, no. 2, pp. 139–157, Aug. 2000.
- [17] C. Domingo and O. Watanabe, "MadaBoost: A modification of adaboost," in *Proc. Annu. Conf. Comput. Learn. Theory (COLT)*, Jun.–Jul. 2000, pp. 180–189.
- [18] P. M. Long and R. A. Servedio, "Random classification noise defeats all convex potential boosters," *Mach. Learn.*, vol. 78, no. 3, pp. 287–304, Mar. 2010.
- [19] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018.
- [20] Z. Wang, R. E. Schapire, and N. Verma, "Error adaptive classifier boosting (EACB): Leveraging data-driven training towards hardware resilience for signal inference," *IEEE Trans. Circuits Syst. I*, vol. 62, no. 4, pp. 1136–1145, Apr. 2015.
- [21] Z. Wang, K. H. Lee, and N. Verma, "Overcoming computational errors in sensing platforms through embedded machine-learning kernels," *IEEE Trans. VLSI Syst.*, vol. 23, no. 8, pp. 1459–1470, Aug. 2015.
- [22] Z. Chair and P. K. Varshney, "Optimal data fusion in multiple sensor detection systems," *IEEE Trans. Aerosp. Electron. Syst.*, vol. AES-22, no. 1, pp. 98–101, Jan. 1986.
- [23] R. Viswanathan and P. K. Varshney, "Distributed detection with multiple sensors: Part I-fundamentals," *Proc. IEEE*, vol. 85, no. 1, pp. 54–63, Jan. 1997.
- [24] X. Li, L. Wang, and E. Sung, "AdaBoost with SVM-based component classifiers," *Eng. Appl. Artif. Intell.*, vol. 21, no. 5, pp. 785–795, Sep. 2008.
- [25] H. Schwenk and Y. Bengio, "Training methods for adaptive boosting of neural networks," in *Proc. Annu. Conf. Neural Inf. Process. Syst. (NIPS)*, Dec. 1997, pp. 647–650.
- [26] ———, "Boosting neural networks," *Neural Comput.*, vol. 12, no. 8, pp. 1869–1887, Aug. 2000.

- [27] B. Chen and P. K. Willett, "On the optimality of the likelihood-ratio test for local sensor decision rules in the presence of nonideal channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 2, pp. 693–699, Feb. 2005.
- [28] B. Chen, L. Tong, and P. K. Varshney, "Channel-aware distributed detection in wireless sensor networks," *IEEE Signal Process. Mag.*, vol. 23, no. 4, pp. 16–26, Jul. 2006.
- [29] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, "On the Lambert W function," *Adv. Comput. Math.*, vol. 5, no. 1, pp. 329–359, Dec. 1996.
- [30] A. Asuncion and D. Newman, "UCI machine learning repository," 2007. [Online]. Available: <http://archive.ics.uci.edu/ml>