Aggregating Binary Classifiers with General Losses

Akshay Balsubramani
UC San Diego
abalsubr@cs.ucsd.edu

Yoav Freund
UC San Diego
yfreund@cs.ucsd.edu

Abstract

We develop a worst-case analysis of aggregation of binary classifier ensembles,for a broad class of losses including but not limited to all convex surrogates. The result is a family of parameter-free ensemble aggregation algorithms, which are as efficient as linear learning and prediction for convex risk minimization but work without any relaxations whatsoever on many nonconvex losses like the 0-1 loss.

1 Introduction

Consider a binary classification problem, in which we attempt to build the best predictor possible for data falling into two classes. At our disposal is an ensemble $\mathcal{H}$ of individual classifiers which we can use in designing our predictor. The task is to predict with minimum error on a large unlabeled test set, on which we know the predictions of the ensemble classifiers but not the true test labels. This is a prototype supervised learning problem, for which a typical solution is to hold out some labeled data to measure the errors of the ensemble classifiers, and then just predict according to the best classifier. But can we use the unlabeled data to better predict using the ensemble classifiers?

This problem was recently studied by the authors [BF15b], who gave a worst-case-optimal algorithm for it when the evaluation metric, and the constraints, are measured with zero-one classification error. However, the zero-one loss is inappropriate for other common binary classification tasks, such as estimating label probabilities, and handling false positives and false negatives differently. Such goals motivate the use of different losses like log loss and cost-weighted misclassification loss.

In this manuscript, we generalize the setup of [BF15b] to these loss functions and others. Like the earlier work, we show that the choice of loss function completely governs an efficient parameter-free aggregation algorithm that predicts with minimax optimally low loss in our setting.

1.1 Preliminaries

Our setting generalizes that of Balsubramani and Freund [BF15b], in which we are given an ensemble $\mathcal{H} = \{h_1, \ldots, h_p\}$ and unlabeled data $x_1, \ldots, x_n$ on which we wish to predict. To start with, the ensemble's predictions on the unlabeled data are denoted by $\mathbf{F}$:

$$\mathbf{F} = \begin{pmatrix} h_1(x_1) & h_1(x_2) & \cdots & h_1(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ h_p(x_1) & h_p(x_2) & \cdots & h_p(x_n) \end{pmatrix} \in [-1,1]^{p \times n} \tag{1}$$

We use vector notation for the rows and columns of $\mathbf{F}$: $\mathbf{h}_i = (h_i(x_1), \cdots, h_i(x_n))^\top$ and $\mathbf{x}_j = (h_1(x_j), \cdots, h_p(x_j))^\top$. The test set has some binary labels $(y_1; \ldots; y_n) \in \{-1,1\}^n$. As in [BF15b], though, the test labels are allowed to be randomized, represented by values in $[-1,1]$ instead of just the two values $\{-1,1\}$. So it is convenient to write the labels on the test data as $\mathbf{z} = (z_1; \ldots; z_n) \in [-1,1]^n$. These true test set labels are unknown to the predictor.

Write $[a]_+ = \max(0, a)$ and $[n] = \{1, 2, \ldots, n\}$. All vector inequalities are componentwise.

1.2 Loss Functions

On any single test point with randomized binary label $z_j \in [-1, 1]$, our expected performance upon predicting g_j, with respect to the randomization of z_j, is measured by a loss function $\ell(z_j, g_j)$. It is apparent that $\ell(z_j, g_j) = \left(\frac{1+z_j}{2}\right)\ell(1, g_j) + \left(\frac{1-z_j}{2}\right)\ell(-1, g_j) := \left(\frac{1+z_j}{2}\right)\ell_+(g_j) + \left(\frac{1-z_j}{2}\right)\ell_-(g_j)$ where we conveniently write $\ell_+(g_j) := \ell(1, g_j)$ and $\ell_-(g_j) := \ell(-1, g_j)$. We call $\ell_\pm$ the *partial losses*, following earlier work [RW10]. In this manuscript, we make an assumption on $\ell_+(\cdot)$ and $\ell_-(\cdot)$:

Assumption 1. *Over the interval $(-1, 1)$, $\ell_+(\cdot)$ is decreasing* [1] *and $\ell_-(\cdot)$ is increasing, and both are twice differentiable.*

We view Assumption 1 as natural, because the loss function intuitively measures discrepancy to the true label ± 1. (Differentiability is convenient for our proofs, but most of our arguments do not require it.) Notably, we do *not* require convexity or symmetry of the losses. We refer to losses satisfying Assumption 1 as "general losses" to contrast them with convex losses or other less broad subclasses.

1.3 Evaluation with General Losses

The idea of [BF15b] is to formulate the ensemble aggregation problem as a two-player zero-sum game between a predictor and an adversary. In this game, the predictor is the first player, who plays $\mathbf{g} = (g_1; g_2; \ldots; g_n)$, a randomized label $g_j \in [-1, 1]$ for each example $\{x_j\}_{j=1}^n$. The adversary then sets the labels $\mathbf{z} \in [-1, 1]^n$.

An error bound on a classifier's predictions can be viewed as a constraint on $\mathbf{z}$. Accordingly, we assume the predictor has knowledge of a *correlation vector* $\mathbf{b} \in (0, 1]^p$ such that $\forall i \in [p]$, $\frac{1}{n}\sum_{j=1}^n h_i(x_j) z_j \geq b_i$, i.e. $\frac{1}{n}\mathbf{F}\mathbf{z} \geq \mathbf{b}$. These p inequalities represent upper bounds on individual classifier zero-one error rates, which can be estimated from the training set w.h.p. when the training and test data are i.i.d. (statistical learning), just as in ERM [BF15b]. [2] So in our game-theoretic formulation, the adversary plays under ensemble classifier error constraints defined by $\mathbf{b}$.

The predictor's goal is to *minimize the worst-case expected loss on the test data* (w.r.t. the randomized labeling $\mathbf{z}$), which we write $\ell(\mathbf{z}, \mathbf{g}) := \frac{1}{n}\sum_{j=1}^n \ell(z_j, g_j)$. The predictor's worst-case goal can be written as the following optimization problem, a game:

$$V := \min_{\mathbf{g} \in [-1,1]^n} \max_{\substack{\mathbf{z} \in [-1,1]^n, \\ \frac{1}{n}\mathbf{F}\mathbf{z} \geq \mathbf{b}}} \ell(\mathbf{z}, \mathbf{g}) \tag{2}$$

$$= \frac{1}{2} \min_{\mathbf{g} \in [-1,1]^n} \max_{\substack{\mathbf{z} \in [-1,1]^n, \\ \frac{1}{n}\mathbf{F}\mathbf{z} \geq \mathbf{b}}} \frac{1}{n}\sum_{j=1}^n \left[\ell_+(g_j) + \ell_-(g_j) + z_j\left(\ell_+(g_j) - \ell_-(g_j)\right)\right] \tag{3}$$

In this manuscript, our goal is to solve the learning problem faced by the predictor, finding an optimal strategy $\mathbf{g}^*$ realizing the minimum in (2). This strategy guarantees good worst-case performance on the unlabeled dataset, with an upper bound of V on the loss. This bound is perfectly tight, by virtue of the argument above; for all $\mathbf{z}_0$ and $\mathbf{g}_0$ obeying the constraints, we see from the definitions that $\min_{\mathbf{g} \in [-1,1]^n} \ell(\mathbf{z}_0, \mathbf{g}) \leq V \leq \max_{\substack{\mathbf{z} \in [-1,1]^n, \\ \frac{1}{n}\mathbf{F}\mathbf{z} \geq \mathbf{b}}} \ell(\mathbf{z}, \mathbf{g}_0)$.

In this work, we give the exact minimax $\mathbf{g}^* \in [-1, 1]^n$ for general losses. This turns out to depend componentwise on a linear combination of the input hypotheses, through a sigmoid nonlinearity reminiscent of a "link function." But learning is just a p-dimensional convex optimization problem (Section 2.2), as we prove for a broad subclass of losses (the conditions of Lemma 5), including all convex ERM surrogate losses. Proofs and further extensions are in the full version [BF15a].

[1] Our basic analysis holds more generally, mutatis mutandis, for functions that are not *strictly* decreasing.

[2] The zero-one loss is linear in $\mathbf{z}$, but so is any other general classification loss in our setup; see [BF15a] for details. So without losing generality, the constraints could be general loss bounds.

2 Results

The loss-based **score function** $\Gamma : [-1, 1] \mapsto \mathbb{R}$ is

$$\Gamma(g) := \ell_-(g) - \ell_+(g)$$

(We will also write the vector $\Gamma(\mathbf{g})$ componentwise with $[\Gamma(\mathbf{g})]_j = \Gamma(g_j)$ for convenience, so that $\Gamma(\mathbf{h}_i) \in \mathbb{R}^n$ and $\Gamma(\mathbf{x}_j) \in \mathbb{R}^p$.) Observe that by our assumptions, $\Gamma(g)$ is increasing on its domain. Therefore, we can discuss its inverse $\Gamma^{-1}(m)$. This can be thought of as a sort of link function.

With these in mind, we can set up the solution to the game (2).

Definition 1 (Potential Well)**.** *Define the **potential well***

$$\Psi(m) = \begin{cases} -m + 2\ell_-(-1) & \text{if } \ m \le \Gamma(-1) \\ \ell_+(\Gamma^{-1}(m)) + \ell_-(\Gamma^{-1}(m)) & \text{if } \ m \in (\Gamma(-1), \Gamma(1)) \\ m + 2\ell_+(1) & \text{if } \ m \ge \Gamma(1) \end{cases} \tag{4}$$

As in [BF15b], we show that $\mathbf{g}^*$ is a simple function of a particular weighting over the p hypotheses – a non-negative p-vector.

Definition 2 (Slack Function)**.** *Let $\sigma \ge 0^p$ be a weight vector over $\mathcal{H}$ (not necessarily a distribution). The vector of **ensemble predictions** is $\mathbf{F}^\top \sigma = (\mathbf{x}_1^\top \sigma, \ldots, \mathbf{x}_n^\top \sigma)$, whose elements' magnitudes are the **margins**. The **prediction slack function** is*

$$\gamma(\sigma, \mathbf{b}) := \gamma(\sigma) := -\mathbf{b}^\top \sigma + \frac{1}{n} \sum_{j=1}^{n} \Psi(\mathbf{x}_j^\top \sigma) \tag{5}$$

*An **optimal weight vector** σ^* is any minimizer of the slack function: $\sigma^* \in \arg\min_{\sigma \ge 0^p} [\gamma(\sigma)]$.*

2.1 Solution of the Game

These are used to describe the minimax equilibrium of the game (2), in our main result.

Theorem 3. *The minimax value of the game* (2) *is*

$$\min_{\mathbf{g} \in [-1,1]^n} \max_{\substack{\mathbf{z} \in [-1,1]^n, \\ \frac{1}{n}\mathbf{F}\mathbf{z} \ge \mathbf{b}}} \ell(\mathbf{z}, \mathbf{g}) = V = \frac{1}{2}\gamma(\sigma^*) = \frac{1}{2} \min_{\sigma \ge 0^p} \left[-\mathbf{b}^\top \sigma + \frac{1}{n} \sum_{j=1}^{n} \Psi(\mathbf{x}_j^\top \sigma) \right]$$

The minimax optimal predictions are defined as follows: for all $i \in [n]$,

$$g_j^* := g_j(\sigma^*) = \begin{cases} -1 & \text{if } \ \mathbf{x}_i^\top \sigma^* \le \Gamma(-1) \\ \Gamma^{-1}(\mathbf{x}_i^\top \sigma^*) & \text{if } \ \mathbf{x}_i^\top \sigma^* \in (\Gamma(-1), \Gamma(1)) \\ 1 & \text{if } \ \mathbf{x}_i^\top \sigma^* \ge \Gamma(1) \end{cases} \tag{6}$$

We can also redo the proof of Theorem 3 when $\mathbf{g} \in [-1, 1]^n$ is not left as a free variable set in the game, but instead is preset to $\mathbf{g}(\sigma)$ as in (6) for some (possibly suboptimal) weight vector σ.

Observation 4. *For any weight vector $\sigma_0 \ge 0^p$, the worst-case loss after playing $\mathbf{g}(\sigma_0)$ is bounded by*

$$\max_{\substack{\mathbf{z} \in [-1,1]^n, \\ \frac{1}{n}\mathbf{F}\mathbf{z} \ge \mathbf{b}}} \ell(\mathbf{z}, \mathbf{g}(\sigma_0)) \le \frac{1}{2}\gamma(\sigma_0)$$

It is also instructive to outline some properties of the potential well (and slack function).

Lemma 5. *The potential well $\Psi(m)$ is continuous and 1-Lipschitz. It is also convex under* any *of the following conditions:*

(A) The partial losses $\ell_\pm(\cdot)$ are convex over $(-1, 1)$.
(B) The loss function $\ell(\cdot, \cdot)$ is a proper loss ([SJAM66, RW10]).
(C) $\ell'_-(x)\ell''_+(x) \ge \ell''_-(x)\ell'_+(x)$ for all $x \in (-1, 1)$.

(Indeed, the proof shows that the last condition is both sufficient and necessary for convexity of Ψ, under Assumption 1.) Note that these conditions encompass convex surrogate losses commonly used in ERM, including all such "margin-based" losses (convex univariate functions of $z_j g_j$). These constitute a large class of losses introduced primarily for their favorable computational properties relative to 0-1 loss ERM.

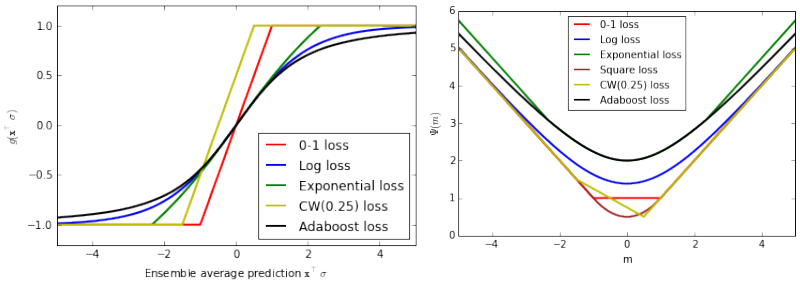

Figure 1: At left are optimal prediction functions g, as a function of margin. At right are plots of the potential wells for various common losses. For more on the losses, see [BF15a].

2.2 The Ensemble Aggregation Algorithm

Theorem 3 defines a prescription for aggregating the given ensemble predictions on the test set. This can be stated in terms of a learning algorithm and a prediction method.

Learning. *Minimize the slack function $\gamma(\sigma)$, finding the minimizer σ^* that achieves V.*
This is a convex optimization under broad conditions (Lemma 5), and when the test examples are i.i.d. the Ψ term is a sum of n i.i.d. functions. As such it is readily amenable even to standard first-order optimization methods like stochastic gradient descent and variants. In practice, learning employs such methods to *approximately* minimize γ, finding some σ_A such that $\gamma(\sigma_A) \leq \gamma(\sigma^*) + \epsilon$ for some small ϵ. Standard convex optimization methods will do this because the slack function is Lipschitz, as Lemma 5 shows (combined with the observation that $\|\mathbf{b}\|_\infty \leq 1$). □

Prediction. *Predict $g(\sigma^*)$ on any test example, as indicated in* (6).
This decouples the prediction task on each test example, which is as efficient as p-dimensional linear prediction, requiring $O(p)$ time and memory. After finding an ϵ-approximate minimizer σ_A in the learning step, Observation 4 tells us that the prediction $\mathbf{g}(\sigma_A)$ has loss within $\frac{\epsilon}{2}$ of V. □

2.3 Discussion and Extensions

The work [BF15b] addresses a problem, 0-1 loss minimization, that is well known to be strongly NP-hard when solved directly. Formulating it in the transductive setting, in which the data distribution is known, is crucial. It gives the dual problem a special meaning, so the learning problem is on the always-convex Lagrange dual function and is therefore tractable. This work generalizes that idea, as the possibly non-convex partial losses are minimized transductively via a straightforward convex optimization. It would be interesting to investigate these new algorithms in practice, given the results for 0-1 loss [BF15c].

Our transductive formulation involves no surrogates or relaxations of the loss, which allows us to bypass the consistency and agnostic-learning discussions [Zha04, BJM06] common to ERM methods that use convex risk minimization. Convergence analyses of such methods make heavy use of convexity of the losses and are generally done presupposing a linear weighting on $h \in \mathcal{H}$ [TDS15], whereas in our work such structure emerges directly from Lagrange duality and involves no convexity to derive the worst-case-optimal predictions. However, prior work does express the conclusion we explicitly derive – the learning problem is completely determined by the choice of loss function.

All our algorithms in this manuscript can be used in full generality with "specialist" hypotheses in the ensemble that only predict on some subset of the test examples. This is done by merely changing $\mathbf{F}$ and $\mathbf{b}$ so that the loss bounds are only over these examples; see [BF15c].

In minimizing the slack function over the dual parameters σ, we perform at least as well as the weighting $\sigma^i \geq \mathbf{0}^p$ that puts weight 1 on h_i and 0 on the remaining classifiers $h_{i' \neq i}$. In other words, our predictor always has the option of simply choosing the best single classifier i^* and guaranteeing its loss bound $\epsilon^\ell_{i^*}$. Consequently, our predictor's loss is always at most that of any single classifier. For the same reason, our algorithm automatically admits superior worst-case loss bounds to *any weighted majority vote* as well, given the ensemble loss constraints $\mathbf{b}^\ell$.

References

[BF15a] Akshay Balsubramani and Yoav Freund. Minimax binary classifier aggregation with general losses. *CoRR*, abs/1510.00452, 2015.

[BF15b] Akshay Balsubramani and Yoav Freund. Optimally combining classifiers using unlabeled data. In *Conference on Learning Theory*, 2015.

[BF15c] Akshay Balsubramani and Yoav Freund. Scalable semi-supervised classifier aggregation. In *Advances in Neural Information Processing Systems*, 2015.

[BJM06] Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

[RW10] Mark D Reid and Robert C Williamson. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

[SJAM66] Emir H Shuford Jr, Arthur Albert, and H Edward Massengill. Admissible probability measurement procedures. *Psychometrika*, 31(2):125–145, 1966.

[TDS15] Matus Telgarsky, Miroslav Dudik, and Robert Schapire. Convex risk minimization and conditional probability estimation. In *Conference on Learning Theory*, 2015.

[Zha04] Tong Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, pages 56–85, 2004.