K-hyperplane Hinge-Minimax Classifier

Margarita Osadchy RITA@CS.HAIFA.AC.IL
Tamir Hazan TAMIR@CS.HAIFA.AC.IL
Daniel Keren DKEREN@CS.HAIFA.AC.IL
Department of Computer Science, University of Haifa, Mount Carmel, Haifa 31905, Israel

Abstract

We explore a novel approach to upper bound the misclassification error for problems with data comprising a small number of positive samples and a large number of negative samples. We assign the hinge-loss to upper bound the misclassification error of the positive examples and use the minimax risk to upper bound the misclassification error with respect to the worst case distribution that generates the negative examples. This approach is computationally appealing since the majority of training examples (belonging to the negative class) are represented by the statistics of their distribution, in contrast to kernel SVM which produces a very large number of support vectors in such settings. We derive empirical risk bounds for linear and non-linear classification and show that they are dimensionally independent and decay as $1/\sqrt{m}$ for m samples. We propose an efficient algorithm for training an intersection of finite number of hyperplanes and demonstrate its effectiveness on real data, including letter and scene recognition.

1. Introduction

Linear classifiers are the cornerstone of several applications in machine learning. The generalization ability of linear classifiers has been long studied in the context of support vector machines (SVMs), e.g., using VC dimension, covering numbers, and Rademacher complexity (Vapnik, 2000; Zhang, 2002; Bartlett & Mendelson, 2003; Bousquet et al., 2004; Kakade et al., 2009). SVMs upper bound the misclassification loss of the linear classifier using the hinge-loss. This setting is computationally appealing when there are fairly small number of support vectors. Alternatively, the minimax risk upper bounds the distribution that generates the instances-labels examples in the world (Lanckriet et al., 2003; Honorio & Jaakkola, 2014). This approach is computationally appealing when there are (infinitely) many training examples since it only utilizes their statistical properties, such as mean and covariance. In our work we consider real-life data that consists of a small number of positive data points and a large number of negative data points, a setting that is prominent in machine learning applications, e.g., in computer vision, such as object detection, and in security, such as fraud or malicious attack detection. We suggest to combine the hinge-risk with the minimax risk to enjoy the best of both worlds. The idea of combining minimax for the negative class and svm-like formulation for the positive samples was introduced in (Osadchy et al., 2012) for training a single hyperplane. No generalization bounds have been shown in that work. Here we derive an empirical mixed-risk bound, that uses the Rademacher complexities to bound the hinge-risk and vector Bernstein's inequalities to bound the minimax risk. Recently, (Honorio & Jaakkola, 2014) derived a generalization bound for the minimax risk using PAC-Bayesian approach, a setting that bounds the expected loss with respect to a posterior distribution over all possible classifiers. Our work differs as we use stronger assumptions - that the norm of the data points is bounded, an assumption that is natural in many applications. Thus we are able to avoid the variance while computing the expected loss of PAC-Bayesian bounds.

Our work mainly considers non-linear classifiers. Non-linear classifiers are usually attained by applying kernel methods. Unfortunately, the computational complexity of kernel methods increases linearly with the size of the training sample (Steinwart, 2003). Instead we apply non-linearity by using intersection of hyperplanes (Klivans & Servedio, 2004; Arriaga & Vempala, 1999). Unfortunately, the proposed algorithms for intersection of hyperplanes are computationally costly when considering large sets of negative data points. To deal with this computational difficulty we extend the minimax risk to deal with intersection of hyperplanes over (infinitely many) negative examples. As this risk bound is loose when there are few

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

data points, as we have in the positive examples, we apply the sum-of-hinge-loss to the positive examples. We also derive a generalization bound that mixes these two risks for intersection of hyperplanes setting. Our Rademacher complexity bound for classifying by intersection of hyperplanes extends the contraction lemma (cf. (Bartlett & Mendelson, 2003)) to vectors. This may be used to simplify recent multi-class generalization bounds using Rademacher complexity (Cortes et al., 2013).

We propose an algorithm for training an intersection of hyperplanes that efficiently minimizes the minimax-hinge risk. We show empirically on two real sets with very differing characteristics, that this algorithm substantially improves over linear classifiers; further, it is on par with the classification rate of ensemble methods (comprising more than a 100 simple classifiers compared to 2-4 hyperplanes in the hinge-minimax classifier) and it even approached the classification performance of kernel SVM, but is orders of magnitude faster.

2. Background

Let $(x, y) \sim D$, where $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$. Let $y_w(x) = sign(w^T x)$ denote a linear classifier. For simplicity, we assume that $b = 0$ (or absorbed by w). A zero-one risk for the $y_w(x)$ is defined as follows:

$$L_D^{0/1}(w) = \mathbb{E}_D[1[y_w(x) \neq y]]$$

The zero-one loss is non-convex. In SVMs the zero-one loss is upper bounded by the hinge loss: $1[y_w(x) \neq y] \leq \max\{0, 1 - yw^T x\}$. Thus the hinge risk upper bounds the zero-one risk:

$$L_D^{0/1}(w) \leq \mathbb{E}_D[\max\{0, 1 - yw^T x\}] \triangleq L_D^H(w).$$

Alternatively, one can upper bound the zero-one risk by the minimax risk (Lanckriet et al., 2003; Honorio & Jaakkola, 2014). There, instead of upper bounding the zero-one loss function, one upper bounds the distribution that generated the data. Denote $\mu = \mathbb{E}_{(x,y)\sim D}[yx]$ and $\Sigma = \mathbb{E}_{(x,y)\sim D}[(yx - \mu)(yx - \mu)^T]$. Denote by $Z(\mu, \Sigma)$ the set of all distributions with mean μ and covariance Σ.

$$L_D^{0/1}(w) \leq \sup_{yx \sim Z(\mu,\Sigma)} Pr(w^T yx \leq 0) \triangleq L_{\mu,\Sigma}^M(w)$$

It was shown in (Lanckriet et al., 2003; Honorio & Jaakkola, 2014) that

$$\sup_{z \sim Z(\mu,\Sigma)} Pr(w^t z \leq 0) = \frac{1}{1 + \frac{(w^T \mu)^2}{w^T \Sigma w}}.$$

In our work, we consider non-linear classification with K hyperplanes. Let w_i, $i = 1, .., K$ denote K hyperplanes. Let W be a matrix with w_i as its ith column. We define a non-linear classifier $f_W(x)$ as an intersection of these K hyperplanes:

$$f_W(x) = \begin{cases} 1 & \text{if } W^T x \geq 0 \\ -1 & \text{otherwise} \end{cases} \quad (1)$$

where 0 denotes a vector of zeros.

3. Minimax-Hinge Risk

We are interested in a classification problem in which the positive class corresponds to a single concept and the negative class is its complement. We assume that the sample of the positive class is relatively small while the negative sample is very large (it can be represented by an unlabeled data as well, thus is easy to collect). When the sample is small, the regularized hinge loss has been shown to be very effective. The minimax bound is tight for the Gaussian distribution, thus it will become tight for sample size approaching infinity. Due to the specifics of the problem we propose to use a mixed risk, namely, we use hinge risk for the positive class and minimax risk for the negative. Next, we formulate the mixed risk for the non-linear classifier in eq. 1.

$$L_D^{MH}(W) = L_D^{H,1} + L_{\mu(D_{neg}),\Sigma(D_{neg})}^{M,-1} \quad (2)$$

where W is the matrix of K hyperplanes, D is a joint distribution of the samples and labels, D_{neg} is a marginal distribution of samples over the negative labels, and $\mu(D_{neg})$ amd $\Sigma(D_{neg})$ are its mean and covariance respectively. $L_D^{H,1}$ and $L_{\mu(D_{neg}),\Sigma(D_{neg})}^{M,-1}$ are defined below.

$$L_D^{H,1} = \mathbb{E}_D[L^H(W, x, y)1[y = 1]]$$

where $L^H(W, x, 1) = \sum_i \max\{0, 1 - w^T x\}$

3.1. The Expected Risk of the Negative Class

Since $L_{\mu(D_{neg}),\Sigma(D_{neg})}^{M,-1}$ is the risk of a negative sample falling into the intersection of K hyperplanes, which is a convex set, we can use the theorem due to Marshall and Olkin (Marshall & Olkin, 1960) to derive the risk for the negative class.

Theorem 1
Let $Z(\mu, \Sigma)$ [1] be all distributions with known mean μ and covariance Σ. For K fixed hypeplanes w_i $(i = 1, .., K)$, we have

$$\sup_{z \sim Z(\mu,\Sigma)} \Pr(W^T z > 0) = \frac{1}{1 + d^2}$$

with $d^2 = \mu^T \tilde{W}(\tilde{W}^T \Sigma \tilde{W})^{-1} \tilde{W}^T \mu$, where $\tilde{W}$ is a matrix with columns that satisfy $w^T z^* = 0$, where

[1] for notational simplicity, we drop D_{neg} and denote $\mu = \mu(D_{neg})$ and $\Sigma = \Sigma(D_{neg})$.

$z^* = \arg\min_z (z-\mu)^T \Sigma^{-1} (z-\mu)$.

Proof. Following the result of Marshall and Olkin (Marshall & Olkin, 1960) for a convex set, we obtain:

$$\sup_{z \sim Z(\mu,\Sigma)} \Pr(W^T z \geq 0) = \frac{1}{1+d^2},$$

with $d^2 = \inf_{W^T z \geq 0} (z-\mu)^T \Sigma^{-1} (z-\mu)$ Next, we want to derive a closed-form expression for d^2. We seek the solution for the primal problem

$$\min_z (z-\mu)^T \Sigma^{-1} (z-\mu)$$

s.t. $w_i^T z \geq 0$ for $i = 1, .., K$. We construct the Lagrangian:

$$L(z, \lambda_i) = (z-\mu)^T \Sigma^{-1} (z-\mu) + \sum_i \lambda_i w_i^T z, \quad \lambda_i \geq 0.$$

The optimality condition:

$$\frac{\partial L}{\partial z} = 2\Sigma^{-1} z - 2\Sigma^{-1}\mu + \sum_i \lambda_i w_i = 0,$$

gives us $z^* = \mu - \frac{1}{2}\sum_i \lambda_i \Sigma w_i$. The Lagrange dual function is as follows,

$$L(z^*, \lambda) = (\frac{1}{2}\sum_i \lambda_i \Sigma w_i)^T \Sigma^{-1} (\frac{1}{2}\sum_j \lambda_j \Sigma w_j) \quad (3)$$
$$+ \sum_i \lambda_i w_i^t (\mu - \frac{1}{2}\sum_j \lambda_j \Sigma w_j)$$

The optimality conditions are:

$$\frac{\partial L(z^*, \lambda)}{\partial \lambda_k} = -\frac{1}{2}\sum_i \lambda_i w_k^T \Sigma w_i + w_k^T \mu = 0$$

for k such that $\lambda_k > 0$.

The function is optimized at

$$\lambda^* = 2(\tilde{W}\Sigma\tilde{W})^{-1}\tilde{W}^T\mu, \quad (4)$$

$\tilde{W}$ is formed by a subset of columns of W for which $\lambda_k > 0$, and thus $w_k^T z^* = 0$.

For the last step we substitute the optimal λ, given in eq. 4 into the dual function in eq. 3 and after simple algebraic manipulations we get:

$$d^2 = \max_{\lambda \geq 0} (L(z^*, \lambda^*)) = \mu^T \tilde{W} (\tilde{W}^T \Sigma \tilde{W})^{-1} \tilde{W}^T \mu$$

□

4. Bound

In the following we bound the risk by its finite sample. In what follows we show that the discrepancy between the risk $L_D^{MH}(W)$ and its empirical estimation $L_S^{MH}(W)$ decays at the rate of $c\sqrt{\frac{\log(1/\delta)}{m}}$ where δ is the confidence over the samples of the training data and m is the training data size. The main difficulty arises from mixing the hinge-risk for the positive examples and the minimax risk for the negative examples. For this purpose we divide the risk to its positive and negative cases, and for each we derive a uniform generalization bound.

4.1. Uniform generalization bound for the empirical minimax risk

The minimax risk upper bounds the zero-one risk over the negative examples. To derive a uniform generalization bound for this setting we bound each of its components differently using a finite sample of training examples. In view of eq. 2 we bound the relative size of the negative set by its empirical average. We also bound the minimax risk itself, by its training sample estimation.

For the sake of clarity we begin with deriving a generalization bound for a single hyperplane, followed by a generalization bound for the intersection of hyperplanes. Recall that D_{neg} is the distribution of the negative data points. We abbreviate its mean and covariance by $\mu \triangleq \mu(D_{neg})$ and $\Sigma \triangleq \Sigma(D_{neg})$ respectively. Let $\hat{\mu}$ and $\hat{\Sigma}$ be the mean and covariance estimates from the training data points that are associated with negative labels. The minimax generalization bound $L_{D(\mu,\Sigma)}^{M,-1}(w)$ is dominated by the discrepancy

$$\Delta = \frac{1}{1+\frac{(w^\top \mu)^2}{w^\top \Sigma w}} - \frac{1}{1+\frac{(w^\top \hat{\mu})^2}{w^\top \hat{\Sigma} w}}$$

Some algebraic manipulations yield a simpler form of the discrepancy:

$$\Delta = \frac{w^\top \Sigma w \cdot (w^\top \hat{\mu})^2 - w^\top \hat{\Sigma} w \cdot (w^\top \mu)^2}{\left(w^\top \Sigma w + (w^\top \mu)^2\right) \cdot \left(w^\top \hat{\Sigma} w + (w^\top \hat{\mu})^2\right)} \quad (5)$$

To provide uniform generalization bound to the minimax risk, we show that the discrepancy Δ decreases when the size of the training sample increases. Therefore we represent the discrepancy with $\|\hat{\mu} - \mu\|$ and $\|\hat{\Sigma} - \Sigma\|$ that decrease as a function of the training sample. By adding and subtracting $(w^\top \mu)^2 \cdot w^\top \Sigma w$ to the numerator we are able to represent the discrepancy with these diminishing quantities. $\Delta =$

$$\frac{w^\top \Sigma w \cdot \left((w^\top \hat{\mu})^2 - (w^\top \mu)^2\right) + (w^\top \mu)^2 \cdot w^\top (\Sigma - \hat{\Sigma}) w}{\left(w^\top \Sigma w + (w^\top \mu)^2\right) \cdot \left(w^\top \hat{\Sigma} w + (w^\top \hat{\mu})^2\right)} \quad (6)$$

The sampled quantity $\|\hat{\mu} - \mu\|$ diminishes with high probability as the training size increases. This controls the first term of the minimax discrepancy, as described by the following lemma:

Lemma 1
Assume $x \sim D_{neg}$ is a distribution over data points x with negative labels such that $\|x\| \leq 1$ holds with probability 1. Denote by μ its mean by Σ its covariance. Let S_{neg} training sample of size $\hat{m}$ and let $\hat{\mu} = \frac{1}{\hat{m}} \sum_{x \in S_{neg}} x$ be its sampled mean and $\hat{\Sigma} = \frac{1}{\hat{m}} \sum_{x \in S_{neg}} (x - \hat{\mu})(x - \hat{\mu})^\top$ be its sampled covariance. Define

$$\Delta_1 = \frac{w^\top \Sigma w \cdot \left((w^\top \hat{\mu})^2 - (w^\top \mu)^2\right)}{\left(w^\top \Sigma w + (w^\top \mu)^2\right) \cdot \left(w^\top \hat{\Sigma} w + (w^\top \hat{\mu})^2\right)}.$$

Assume that the minimal eigenvalue of $\Sigma, \hat{\Sigma}$ is lower bounded by α. Then, with probability at least $1 - \delta$ over the draws of the training set S_{neg} the following holds uniformly for all w

$$\Delta_1 \leq \frac{2}{\alpha} \sqrt{\frac{32 \log(1/\delta) + 1/4}{\hat{m}}}$$

Proof. First, we upper bound Δ_1 while decreasing the denominator, by omitting $(w^\top \mu)^2$ and $(w^\top \hat{\mu})^2$. Using the identity $a^2 - b^2 = (a + b)(a - b)$ with $a = w^\top \hat{\mu}$ and $b = w^\top \mu$ we obtain:

$$\Delta_1 \leq \frac{w^\top (\hat{\mu} - \mu) \cdot w^\top (\hat{\mu} + \mu)}{w^\top \hat{\Sigma} w}.$$

Next, we applying the Cauchy-Schwarz inequality to the numerator $a^\top b \leq \|a\| \|b\|$ and the lower bound $w^\top \hat{\Sigma} w \geq \alpha \|w\|^2$ to the denominator

$$\Delta_1 \leq \frac{\|\hat{\mu} - \mu\| \cdot \|\hat{\mu} + \mu\|}{\alpha}.$$

Finally, since $\|x\| \leq 1$ then $\|\hat{\mu} + \mu\| \leq 2$. Moreover, Bernstein inequality for vectors (cf. (Gross, 2011) Theorem 11, (Candes & Plan, 2011) Theorem 2.6) implies $P[\|\hat{\mu} - \mu\| \geq t] \leq \exp(-\hat{m} t^2/32 + 1/4)$ for any $t \leq 2\hat{m}$. The result follows when setting $\lambda = \exp(-\hat{m} t^2/32 + 1/4)$, or equivalently $t = \sqrt{\frac{32(\log(1/\delta) + 1/4)}{\hat{m}}}$. □

We turn to handle the second term of the minimax discrepancy in eq. 5. The quantity $\|\hat{\Sigma} - \Sigma\|$ diminishes in high probability as the training size increases.

Lemma 2
Under the conditions of Lemma 1, define

$$\Delta_2 = \frac{(w^\top \mu)^2 \cdot w^\top (\Sigma - \hat{\Sigma}) w}{\left(w^\top \Sigma w + (w^\top \mu)^2\right) \cdot \left(w^\top \hat{\Sigma} w + (w^\top \hat{\mu})^2\right)}.$$

Assume that the minimal eigenvalues of $\Sigma, \hat{\Sigma}$ are lower bounded by α. Then, with probability at least $1 - \delta$ over the draws of the training set S_{neg} the following holds uniformly for all w

$$\Delta_2 \leq \frac{1}{\alpha^2} \sqrt{\frac{128(\log(1/\delta) + 1/4)}{\hat{m}}}$$

Proof. First, we lower bound the denominator by $\alpha^2 \|w\|^2$ (when omitting $(w^\top \mu)^2$ and $(w^\top \hat{\mu})^2$) thus upper bounding Δ_2. Noting the $w^\top A w = a^\top b$ where a is a vectorization of A and b is a vectorization of $ww^\top$, we use the Cauchy-Schwarz inequality to upper bound the numerator by $\|\mu\|^2 \|w\|^2 \cdot \|a\| \|b\|$. Since $\|b\| = \|w\|^2$ and $\|\mu\| \leq 1$ we obtain the bound

$$\Delta_2 \leq \|\Sigma - \hat{\Sigma}\| / \alpha^2.$$

We consider the norm $\|\Sigma - \hat{\Sigma}\|$ as the norm of its vectorized form. Using the Bernstein inequality for vectors we obtain that $P[\|\Sigma - \hat{\Sigma}\| \geq t] \leq \exp(-m t^2/128 + 1/4)$ for any $t \leq 4\hat{m}$. The result follows when setting $\delta = \exp(-m t^2/128 + 1/4)$ or equivalently $t = \sqrt{\frac{128(\log(1/\delta) + 1/4)}{\hat{m}}}$. □

Bounds on the discrepancy between the minimax risk and its empirical risk that are uniform for any w guarantee generalization. The above lemmas suggest that the penalty of observing a finite sample space decreases as $1/\hat{m}$. This is summarized in the following theorem.

Theorem 2
Suppose that D is a distribution over $X \times Y$ such that $Y = \{-1, +1\}$ and $X = \{x : \|x\| \leq 1\}$. Let $L^{M,-1}_{\mu,\Sigma}(w)$ be the minimax risk over the negative labels, where μ, Σ are the mean and covariance of the marginal distribution of x over the negative labels. Consider a training sample S of size m which $\hat{m}$ of them have negative label and let $L_{S_{neg}}(w)$ be the empirical minimax risk over the negative labels

$$L_{S_{neg}}(w) = \frac{\hat{m}}{m} \cdot \sup_{z \in Z(\hat{\Sigma}, \hat{\mu})} P[w^\top z \geq 0]$$

where $\hat{\mu}, \hat{\Sigma}$ are the empirical mean and covariance estimation of the marginal distribution of x over the negative training labels. Then, for any $\delta \in (, 1]$ with probability at least $1 - 3\delta$ over the i.i.d. sample of size m holds $L^{M,-1}_{\mu,\Sigma}(w) \leq$

$$L^{M,-1}_{\hat{\mu},\hat{\Sigma}}(w) + c_1 \sqrt{\frac{\log(1/\delta)}{\hat{m}}} + c_2 \sqrt{\frac{\log(1/\delta) + 1/4}{\hat{m}}}.$$

$c_1 = 1/\sqrt{2}$ and $c_2 = \sqrt{128}(1/\alpha + 1/\alpha^2)$ and α is the minimal eigenvalue of $\Sigma, \hat{\Sigma}$.

Proof. We estimate $\rho = E_{(x,y)\sim D}\Big[1[y = -1]\Big]$ by its empirical mean $\hat{m}/m$. From the Hoeffding inequality, $P[\frac{\hat{m}}{m} - \rho \geq t] \leq \exp(-2mt^2)$. Setting $\delta = \exp(-2mt^2)$ we derive the confidence interval $t = \sqrt{\log(1/\delta)/2m}$. Thus combining the above lemmas, with error probability of 3δ the minimax risk is upper bounded by

$$\Big(\frac{\hat{m}}{m} + \sqrt{\frac{\log(1/\delta)}{2m}}\Big)\Big(\frac{1}{1 + \frac{(w^\top \hat{\mu})^2}{w^\top \hat{\Sigma} w}} + \sqrt{\frac{c}{\hat{m}}}\Big).$$

$c = 2\sqrt{32(\log(1/\delta) + 1/4)} + \sqrt{128(\log(1/\delta) + 1/4)}$. We conclude the proof by using $\hat{m}/m \leq 1$ and $\frac{1}{1+\frac{(w^\top \hat{\mu})^2}{w^\top \hat{\Sigma} w}} \leq 1$. □

The same type of bound holds for classification by any finite number of hyperplanes, i.e., $W^\top x \geq 0$. We omit its derivation as it is tedious and follows the same derivations as above.

Theorem 3
Consider the setting of Theorem 2 and let

$$L^{M,-1}_{\mu,\Sigma}(W) = E_{(x,y)\sim D}\Big[1[y = -1]\Big] \cdot \sup_{z \in Z(\Sigma,\mu)} P[W^\top z \geq 0]$$

where W is a matrix whose columns consist of k different hyperplanes. Then, for any $\delta \in (0, 1]$ with probability at least $1 - \delta$ over the i.i.d. sample of size m there holds uniformly for all W:

$$L^{M,-1}_{\mu,\Sigma}(W) \leq L^{M,-1}_{\hat{\mu},\hat{\Sigma}}(W) + c\sqrt{\frac{\log(1/\delta) + 1/4}{\hat{m}}}.$$

$c = 1/\sqrt{2} + \sqrt{128}k(1/\alpha + 1/\alpha^2)$ and α is the minimal eigenvalue of $\Sigma, \hat{\Sigma}$.

In our setting, it is important to assume that the eigenvalues of Σ are lower bounded by α. This assumption implies that the eigenvalues of $\hat{\Sigma}$ are also bounded from below whenever $\hat{m} \gg d$. Using Cauchy-Schwartz inequality for any w of a unit norm there holds $|w^\top \Sigma w - w^\top \hat{\Sigma} w| \leq \|\Sigma - \hat{\Sigma}\|$. Using the Bernstein inequality for vectors we obtain that $P[\|\Sigma - \hat{\Sigma}\| \geq t] \leq \exp(-mt^2/128 + 1/4)$ for any $t \leq 4\hat{m}$. Thus with high probability (that decays exponentially with $\hat{m}$) we obtain that for any w of unit norm holds $w^\top \hat{\Sigma} w \geq w^\top \Sigma w - t$. Since the eigenvalues of Σ are lower bounded by α we have that $w^\top \Sigma w \geq \alpha$ then the eigenvalues of $\hat{\Sigma}$ are also lower bounded away from zero whenever $t < \alpha$.

4.2. Uniform generalization bound for the empirical risk of the hinge-loss

The risk of the hinge-loss upper bounds the zero-one risk over the positive examples. Using Rademacher complexities we derive a uniform generalization bound for $k-$hyperplanes classification $W^\top x \geq 0$, i.e., $w_i^\top x \geq 0$ for each $i = 1, ..., k$. The Rademacher complexity of a bounded set $A \subset R^k$ is

$$R(A) = \frac{1}{m} E_\sigma[\max_{a \in A} \sum_{i=1}^{m} \sigma_i a_i],$$

while $\sigma_i \in \{-1, +1\}$ are i.i.d. and equally probable random variables. The set A describes the loss of the predictors W over a training sample $(x_1, y_1), ..., (x_m, y_m)$, namely $A = \{L(W^\top x_j, y_j)\}_{j=1,...,m}$. Whenever the loss is Lipschitz with respect to $k-$hyperplanes predictions, the Rademacher complexity is bounded by $\sqrt{k/m}$.

Theorem 4
Consider a $k-$hyperplanes loss function $L(W^\top x, y)$ for which each hyperplane satisfies $\|w_i\| \leq 1$ and each data point satisfies $\|x\| \leq 1$. Assume that the loss is Lipschitz for every y, i.e., $|L(\alpha, y) - L(\beta, y)| \leq \sum_{i=1}^{k} |\alpha_i - \beta_i|$. Then its Rademacher complexity is bounded by $R(\{L(W^\top x_j, y_j)\}_{j=1,...,m}) \leq \sqrt{k/m}$.

Proof. First we prove the decompositional lemma (cf. (Kakade et al., 2009)) for the $k-$hyperplane setting:

$$R(\{L(W^\top x_j, y_j)\}_{j=1,...,m}) \leq R(\{w_i^\top x_j\}_{i=1,...,k, j=1,...,m}).$$

For notational convenience, we prove it for $m = 1$ and the general case follows by induction over m. By definition, $R(W, x_1, y_1) = E_\sigma[\max_W \sigma L(W, x_1, y_1)]$ and for simplicity, we denote this Rademacher complexity by R. Since $P[\sigma = -1] = P[\sigma = 1] = 0.5$ there holds, $R = 0.5 \max_W L(W, x_1, y_1) + 0.5 \max_W -L(W, x_1, y_1)$. By duplicating the hyperplanes to W, W' we are able to maximize both cases jointly, $R = 0.5 * \max_{W,\hat{W}} \Big(L(W, x_1, y_1) - L(\hat{W}, x_1, y_1)]\Big)$. By the Lipschitz condition $L(W, x_1, y_1) - L(\hat{W}, x_1, y_1) \leq \sum_i |w_i^\top x_1 - \hat{w}_i^\top x_1|$. Since w_i and $-w_i$ have the same norm, taking the maximum over $\|w_i\| \leq 1$ may generate the absolute value, therefore

$$\forall \hat{\sigma}_i \in \{-1, 1\} \;\; R \leq 0.5 \max_{W,\hat{W}} \Big(\sum_i \hat{\sigma}_i (w_i^\top x_1 - \hat{w}_i^\top x_1)\Big).$$

Thus we are able to take the expectation with respect to $\hat{\sigma}_i$ while maintaining the inequality. Also, we separate the two maximizations while noting that $\hat{\sigma}_i$ and $-\hat{\sigma}_i$ have the same distribution to obtain the result:

$$R \leq E_{\hat{\sigma}} \max_W \Big(\sum_i \hat{\sigma}_i w_i^\top x_1\Big).$$

For general m we get by induction:

$$mR(L(W, x_j, y_j)) \leq E_{\sigma_{i,j}} \max_{\|w_i\| \leq 1} \Big(\sum_{i,j} \sigma_{i,j} w_i^\top x_j).$$

Algorithm 1 Intersection of K hyperplanes classifier

Input: $\{x_i\}, i = 1, .., N_p$ a set of positive examples; $\{z_i\}, i = 1, .., N_u$ a set of negative examples.
Output: W (K hyperplanes)
{The initial greedy step}
Estimate μ and Σ using $\{z_i\}_i^{N_u}$
Find w_1 using eq.8 with μ and Σ.
for k=2 to K **do**
  Estimate μ_k and Σ_k using $\{z_i \mid w_j^T z_i > 0, j = 1, .., k-1\}$
  Find w_k using eq.8 with μ_k and Σ_k.
end for
{The refinement iterations}
Let P_k be the probability $Pr(W^T z > 0)$ in iteration k
while $(P_{k-1} - P_k > \epsilon)$ **do**
  Estimate μ_k and Σ_k using $\{z_i \mid w_j^T z_i > 0, j = 1, .., K; j \neq k\}$.
  Find w_k using eq.8 with μ_k and Σ_k.
end while

Standard Rademacher type arguments, e.g. (Bartlett & Mendelson, 2003) derive the bound $mR(L(W, x_j, y_j)) \leq \sqrt{km}$. □

The above theorem generalizes the standard decomposition lemma (also known as the contraction lemma) to $k-$hyperplanes with any Lipschitz loss. In our setting we consider the sum of hinge loss functions over the positive examples

$$L(W, x, y) = 1[y = 1] \cdot \sum_{i=1}^{k} \max\{0, 1 - w_i^\top xy\}$$

Since this function satisfies the conditions of the theorem above, we may use the standard Rademacher uniform generalization bound. Let $L_D^{H,1}(W) = E_{(x,y)\sim D}\Big[L(W, x, y)\Big]$ be the risk, and let $L_S^{H,1}(W) = \frac{1}{m}\sum_{i=1}^{m} L(W, x_i, y_i)$ be the empirical risk over a training sample of size m. Then, for any $\delta \in (0, 1]$ with probability at least $1 - \delta$ over the i.i.d. sample of size m there holds simultaneously for all $\|w_1\|, ..., \|w_k\| \leq 1$ whenever $\|x\| \leq 1$:

$$L_D^{H,1}(W) \leq L_S^{H,1}(W) + \sqrt{\frac{4k}{m}} + \sqrt{\frac{2\log(2/\delta)}{m}}$$

5. Algorithm

We aim to minimize the risk in eq. 2. To this end, we minimize the empirical risk regularized by the sum of L_2 norms of the K hyperplanes:

$$\frac{C}{2}\sum_i \|w_i\|^2 + L_S^{M,1}(W) + L_S^{H,-1}(W) \qquad (7)$$

This empirical risk is a non convex and non smooth function, hence a gradient based optimization of it is difficult. However, for a single hyperplane, we can write an equivalent optimization problem:

$$\begin{aligned} \underset{w}{\text{minimize}} \quad & \frac{C}{2}\|w\|^2 + \sum_i \max\{0, 1 - w^T x_i\} \\ \text{subject to} \quad & \gamma\sqrt{w^T\Sigma w} + w^T\mu \leq 0. \end{aligned} \qquad (8)$$

where γ is a constant, controlling the probability in the positive half space, namely $\gamma \triangleq \sqrt{2}\mathsf{erf}^{-1}(1 - 2\delta)$ [2], where $\Pr(w^T z) \leq \delta$. Since we seek to minimize this probability, we can assume that $\delta < 1/2$, and thus $\gamma > 0$. The constraint in the optimization problem 8 is convex for $\gamma > 0$, and then the entire problem in 8 is convex.

We propose an iterative algorithm (Algorithm 1) that approximates the solution to the problem in eq. 7. It starts by training K hyperplanes in a greedy manner and then iteratively adjusts each hyperplanes to minimize the $Pr(W^T z \geq 0)$.

Lemma 3
Algorithm 1 minimizes $Pr(W^T z \geq 0)$ at each iteration.

Proof. We show the proof for two hyperplanes; the same proof holds for K hyperplanes. Let $W = [w_1 w_2]$. We can write

$$\begin{aligned} &Pr(W^T z \geq 0) = Pr(w_1^T z \geq 0) \bigwedge Pr(w_2^T z \geq 0) \\ &= Pr(w_1^T z \geq 0) Pr(w_2^T z \geq 0 | w_1^T z \geq 0) \end{aligned}$$

First, w_1 optimizes $Pr(w_1^T z \geq 0)$. Second, the optimization in eq.8 seeks a w_2 that minimizes $Pr(w_2^T z \geq 0 | w_1^T z \geq 0)$.

Let w_1^i and w_2^i be the two hyperplanes after i iterations and $\alpha = Pr({w_1^i}^T z \geq 0) \bigwedge Pr({w_2^i}^T z \geq 0)$ be the current probability of the negative error in the intersection. The algorithm seeks a hyperplane w_1^{i+1} that minimizes $Pr({w_1^{i+1}}^T z \geq 0 | {w_2^i}^T z \geq 0)$, thus

$$\begin{aligned} &Pr({w_1^{i+1}}^T z \geq 0) \bigwedge Pr({w_2^i}^T z \geq 0) \\ &= Pr({w_1^{i+1}}^T z \geq 0 | {w_2^i}^T z \geq 0) Pr({w_2^i}^T z \geq 0) \leq \alpha. \end{aligned}$$

Therefor Algorithm 1 decreases the $Pr(W^T z \geq 0)$ in each iteration. □

The empirical risk of the intersection of K hyperplanes is the sum of hinge losses. In each iteration the algorithm minimizes the hinge loss of one hyperlane, while keeping the rest fixed, thus the algorithm decreases the hinge risk at

[2] the supremum of the minimax risk is attained for the Gaussian distribution

each iteration. Since the empirical risk is the sum of hinge and minimax risks, it follows from Lemma 1 and the above discussion that Algorithm 1 minimizes the empirical risk in each iteration and thus converges.

6. Experiments

To test the proposed K-hyperplane hinge-minimax classifier, we ran experiments in three different scenarios: synthetic 2D data, letter recognition, and large scale scene classification.

During classification, the K-hyperplane classifier incurs only K times the computational complexity of a linear classifier (just K inner products), hence its "natural competitors" are linear classifiers, and we choose linear SVM for the benchmark.

We have also compared the hinge-minimax classifiers to kernel SVM and ensemble-based methods, which incur far longer running times (this is especially true for kernel SVM). The classification rates of the hinge-minimax classifier in all our experiments were comparable to ensemble classifiers which required 100-170 basic classifiers in order to reach similar performance. In experiments with high-dimensional data, the hinge-minimax classifiers preformed as well as kernel SVM.

The SVM classifiers were trained using C-SVC in LIBSVM [3]. We used the CVX optimization package [4] to find a single hyperplane in Algorithm 1. The ensemble classifiers were trained using the Matlab Statistic toolbox.

6.1. Synthetic Data Example

We construct the hinge-minimax classifier for 2D data to illustrate Algorithm 1. We samples 5000 data points from two highly overlapping Gaussians (see Figure 1) with varying ratio of positive (shown in red) and negative (shown in blue) examples. Each class was equally partitioned into training, validation, and test sets. We estimated the mean and covariance from the training data and tuned the parameters (C and γ) and the bias using the validation set. Table 1 shows the AUC for the different ratios of positive and negative examples using an intersection of 5 hyperplanes. These results demonstrate the robustness of the algorithm to unbalanced sets.

Positive fraction	0.01	0.1	0.2	0.3	0.4	0.5
AUC	94.68	94.91	95.07	94.96	94.89	95.83

Table 1. AUC for different size partitions of positive and negative classes

[3] $http://www.csie.ntu.edu.tw/cjlin/libsvm/$
[4] $http://cvxr.com/cvx/download/$

The first five plots in Figure 1 show the result of the initial greedy step for the first, second, third, forth, and fifth hyperplanes respectively. The contour lines in Figure 1 illustrate the covariance of the negative distribution inside the intersection, which is used to find the optimal separation hyperplane, depicted in black. The last plot in Figure 1 shows the final classifier after 25 iterations. It illustrates that the approximation algorithm succeeds in separating the positive set from the background, and that the refinement iterations improve the separation boundary.

6.2. Letter Recognition

These tests were performed on a data set of letters from the UCI Machine Learning Repository (Murphy & Aha, 1994), which includes 16-dimensional feature vectors for the 26 letters in the English alphabet. The letter images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce 20,000 samples. For each letter, we used 100 samples for training, 250 for validation, and the rest for test (about 400 samples per letter). The parameters of all methods have been chosen using the validation set. Since the test set includes 25 times more negatives than positives, which leads to about 96% classification rate by just classifying all inputs as negative, we used EER as a more faithful measure of performance. Table 2 shows the classification rate at EER , averaged over 26 letters, and the average classification times of the tested classifiers.

Method	Classification rate at EER	Classification time
hinge-minimax $K = 1$	89.32	5.6e-07
hinge-minimax $K = 2$	92.98	1.4e-06
hinge-minimax $K = 3$	93.93	1.5e-06
hinge-minimax $K = 4$	94.48	1.7e-06
Linear SVM	84.87	4.6e-07
RBF kernel SVM	96.47	1.7e-03
AdaBoost	92.26	1.0e-03

Table 2. Letter experiments. K corresponds to the number of hyperplanes used in the hinge-minimax classifier. The times are in sec. AdaBoost uses 100 decision trees.

The hinge-minimax classifiers improves over the linear SVM for all K, and for $K > 1$ outperforms Adaboost with much shorter classification time. For this data set, kernel SVM outperforms all methods. However, the 4-hyperplane hinge-minimax classifier comes fairly close to the performance of the kernel SVM, while its classification time is three magnitudes faster.

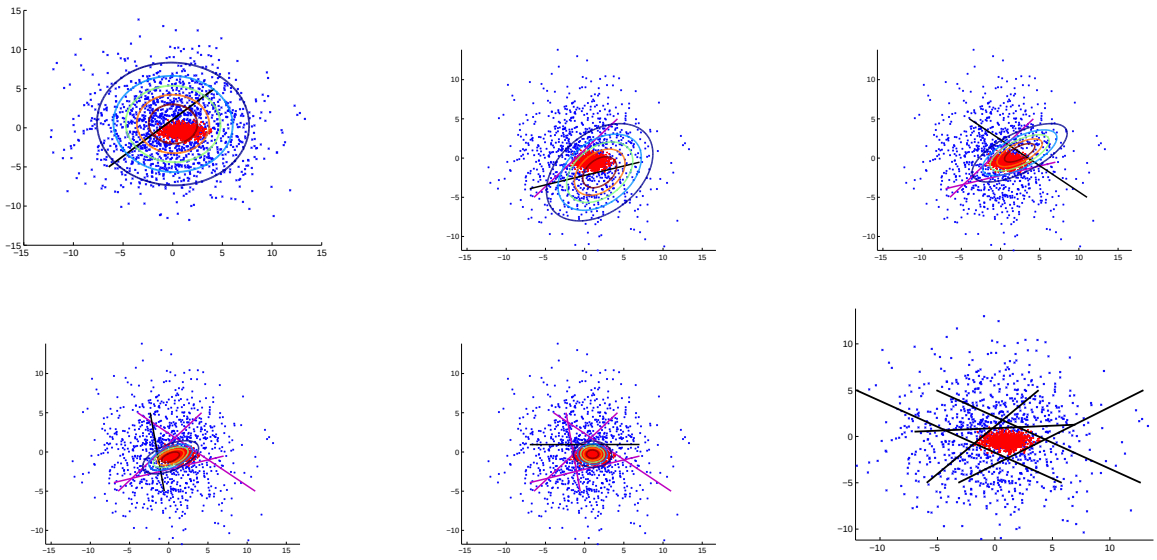

Figure 1. Illustration of hinge-minimax classifier construction on a toy example.The first 5 figures show the greedy initial step. The last figure shows the final classifier after 25 iterations. The contour lines show the covariance matrix of the negative distribution inside the intersection of hyperplane, which is used to find the optimal hyperplane, depicted in black.

6.3. Large Scale Scene Recognition

In this test we used 397 scene categories of the SUN data base, which have at least 100 images per category (Xiao et al., 2010). We represent the images as BOW of dense HOG features with 300 words. We downloaded the features from the SUN web page[5], containing spatial pyramid of BOWs, and used the bottom layer (the details of the feature extraction can be found in (Xiao et al., 2010)). The data is divided into 50 training and 50 test images in 10 folds. Training one-against-all classifiers for 397 categories with 50 training samples per category uses very unbalanced training sets. Thus we defined different weights for positive and negative samples in SVM training and we used RUSBoost (Seiffert et al., 2008) as an ensemble method (it is designed for skewed data and performed significantly better than AdaBoost on this data set). Note that the hinge-minimax classifier naturally handles unbalanced sets. Hinge-minimax classifier with more than two hyperplanes didn't improve the performance. Table 6.3 shows the average AUC which was used for evaluation in (Xiao et al., 2010)) of the tested method and their running times.

Method	AUC	classification time
hinge-minimax, $K = 1$	88.89	9.8e-05
hinge-minimax, $K = 2$	90.99	1.34e-04
Linear SVM	88.20	8.6e-05
RBF kernel SVM	90.77	23.97
RUSBoost	90.76	0.08

Table 3. Scene classification with 300 dim. features. The classification time of RBF kernel SVM is very high, since it chooses about 15,000 SVs from 19850 training examples. The RUSBoost uses 100 decision trees.

7. Conclusions and Future Work

We proposed an efficient method for learning an intersection of finite number of hyperplanes which combined the hinge-risk (for the small number of positive data) with the minimax risk (for a large number of negative data points) and derived a generalization bound for the mixed risk. We show that the proposed classifier yields results comparable to the popular non-linear classifiers, but at much lower (order of magnitude) computational cost of classification.

Extension of this approach to multi-class learning for K hyperplanes remains open. A one-vs-all heuristic is not directly applicable since K-hyperplane intersection yields binary output with no score. One can use the result of Theorem 1 to find the closest distance to the intersection and use it as a score. Another direction is to combine structured-hinge and minimax risks.

Acknowledgments: This work has been supported by Israel Science Foundation 839/12 and the European Union's Seventh Framework Programme FP7-ICT-2013-11 under grant agreement No 619435.

[5] $http://vision.cs.princeton.edu/projects/2010/SUN/$

References

Arriaga, Rosa I and Vempala, Santosh. An algorithmic theory of learning: Robust concepts and random projection. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pp. 616–623. IEEE, 1999.

Bartlett, Peter L and Mendelson, Shahar. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3: 463–482, 2003.

Bousquet, Olivier, Boucheron, Stéphane, and Lugosi, Gábor. Introduction to statistical learning theory. In *Advanced Lectures on Machine Learning*, pp. 169–207. Springer, 2004.

Candes, Emmanuel J and Plan, Yaniv. A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

Cortes, Corinna, Mohri, Mehryar, and Rostamizadeh, Afshin. Multi-class classification with maximum margin multiple kernel. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 46–54, 2013.

Gross, David. Recovering low-rank matrices from few coefficients in any basis. *Information Theory, IEEE Transactions on*, 57(3):1548–1566, 2011.

Honorio, Jean and Jaakkola, Tommi. {Tight Bounds for the Expected Risk of Linear Classifiers and PAC-Bayes Finite-Sample Guarantees}. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 384–392, 2014.

Kakade, Sham M, Sridharan, Karthik, and Tewari, Ambuj. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in neural information processing systems*, pp. 793–800, 2009.

Klivans, Adam R and Servedio, Rocco A. Learning intersections of halfspaces with a margin. In *Learning Theory*, pp. 348–362. Springer, 2004.

Lanckriet, Gert R.G., Ghaoui, Laurent El, Bhattacharyya, Chiranjib, and Jordan, Michael I. A robust minimax approach to classification. *J. Mach. Learn. Res.*, 3:555–582, 2003.

Marshall, Albert W. and Olkin, Ingram. Multivariate chebyshev inequalities. *Ann. Math. Statist.*, 31(4):1001–1014, 1960.

Murphy, P. and Aha, D. Uci repository of machine learning databases. *Tech. rep., U. California, Dept. of Information and Computer Science*, 1994.

Osadchy, M., Keren, D., and Fadida-Specktor, B. Hybrid classifiers for object classification with a rich background. In *ECCV (5)*, pp. 284–297, 2012.

Seiffert, Chris, Khoshgoftaar, Taghi M., Hulse, Jason Van, and Napolitano, Amri. Rusboost: Improving classification performance when training data is skewed. In *ICPR*, pp. 1–4, 2008.

Steinwart, Ingo. Sparseness of support vector machines. *Journal of Machine Learning Research*, 4:1071–1105, 2003.

Vapnik, Vladimir. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., and Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. *CVPR*, pp. 3485–3492, 2010.

Zhang, Tong. Covering number bounds of certain regularized linear function classes. *The Journal of Machine Learning Research*, 2:527–550, 2002.