# Tight Bounds for the Expected Risk of Linear Classifiers and PAC-Bayes Finite-Sample Guarantees

**Jean Honorio**
CSAIL, MIT
Cambridge, MA 02139, USA
jhonorio@csail.mit.edu

**Tommi Jaakkola**
CSAIL, MIT
Cambridge, MA 02139, USA
tommi@csail.mit.edu

## Abstract

We analyze the expected risk of linear classifiers for a fixed weight vector in the "minimax" setting. That is, we analyze the worst-case risk among all data distributions with a given mean and covariance. We provide a simpler proof of the tight polynomial-tail bound for general random variables. For sub-Gaussian random variables, we derive a novel tight exponential-tail bound. We also provide new PAC-Bayes finite-sample guarantees when training data is available. Our "minimax" generalization bounds are dimensionality-independent and $\mathcal{O}(\sqrt{1/m})$ for $m$ samples.

## 1 Introduction

Linear classifiers are the cornerstone of several applications in machine learning. The generalization ability of classifiers have been long studied both in the statistical and computational learning theory. Several general frameworks have been applied in order to analyze the generalization error of generic classifiers (some do not apply to linear classifiers), such as empirical risk minimization, structural risk minimization, VC dimension, covering numbers and Rademacher complexity. Additionally, several notions of stability that guarantee generalization were introduced in [5, 18, 20, 21]. We refer the interested reader to the survey article [4] for additional information.

For linear classifiers, sharp margin bounds are provided in [2] using Rademacher complexity, and [15, 17] using the PAC-Bayes theorem. Later, [11] provided sharp bounds for Rademacher and Gaussian complexities of (constrained) linear classes, which led to several generalization bounds, such as margin bounds, PAC-Bayes bounds and risk bounds for vectors with bounded norm. More recently, [8] generalized the KL divergence in the PAC-Bayes theorem to arbitrary convex functions; and [10] provided dimensionality-dependent PAC-Bayes margin bounds. In a different line of work, [19] proved generalization of sparse linear classifiers (under $\ell_1$-regularization) by using covering number bounds in [23].

Let $\widehat{f}$ be a classifier learnt from $m$ available training samples (drawn from some arbitrary data distribution). Let $f^*$ be the optimal classifier in the asymptotic setting (intuitively speaking, when infinite amount of data is available). Generalization bounds are usually termed as a uniform convergence statement that holds for all classifiers (See [11] for instance). Alternatively, by using a symmetrization argument and by optimality of the empirical minimizer, we can state that with probability at least $1-\delta$:

$$\mathcal{R}(\widehat{f}) \leq \mathcal{R}(f^*) + g(m,\delta) \qquad (1)$$

where $\mathcal{R}(f)$ is the expected risk of the classifier $f$ with respect to the data distribution (and with respect to the posterior for PAC-Bayes), and $g(m,\delta)$ is a function that decreases with respect to $m$ and $\delta$. Very often, we have that $g(m,\delta) \in \mathcal{O}(\sqrt{1/m})$ and $g(m,\delta) \in \mathcal{O}(\sqrt{\log 1/\delta})$ but other rates are also possible [4]. Generalization bounds are stated as in eq.(1) with respect to the unknown quantity $\mathcal{R}(f^*)$. In this paper, we are interested in developing a tight bound for the expected risk $\mathcal{R}(f)$ of a linear classifier $f$. This allows us to find a closed form expression for the bound of $\mathcal{R}(f^*)$ and also study the behavior of $\mathcal{R}(\widehat{f})$ when training data is available.

We study the expected risk of linear classifiers under two scenarios: general random variables and sub-Gaussian variates. Many features used in real-world classification problems follow sub-Gaussianity assump-

Appearing in Proceedings of the $17^{th}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

tions. For instance, in the computer vision literature, histogram features are used for object classification as in [6, 22]. The class of sub-Gaussian variates includes for instance Gaussian variables, any bounded random variable (e.g. Bernoulli, multinomial, uniform), any random variable with strictly log-concave density, and any finite mixture of sub-Gaussian variables. Stronger assumptions have been previously used for the analysis of linear classifiers. In the generalization bound analysis of [11], the authors assumed boundedness; while in the active learning analysis of [1] the authors assumed a log-concave distribution.

In this paper, we provide tight bounds for the expected risk of linear classifiers. Interestingly, the Fisher linear discriminant objective function appears in the different scenarios under our analysis, although our results apply to any linear classifier. In our tightness proof, we construct a family of data distributions where the expected risk bound holds with equality. Our constructions do not rely on conditional distributions of $\mathbf{x}$ given the class $y$ (i.e. $P(\mathbf{x}|y=-1)$ and $P(\mathbf{x}|y=+1)$) that are Gaussian or that have equal covariances. This implies that there is a whole family of non-trivial distributions in which the Fisher discriminant is asymptotically as good as any other linear classifier.

When training data is available, we provide novel "minimax" generalization bounds that are dimensionality-independent and $\mathcal{O}(\sqrt{1/m})$ for $m$ samples. From a technical point of view, we do not require boundedness of the input data as in [11], but note that we analyze a new problem. As part of our analysis, we derive novel PAC-Bayes bounds for *not-everywhere-bounded* functions. Additionally, while PAC-Bayes bounds usually depend on the Kullback-Leibler divergence (which is natural for our analysis of sub-Gaussian variates), we provide bounds that depend on the Chi-squared divergence (which we believe is natural for our analysis of general random variables).

## 2 Preliminaries

Consider a binary classification problem with $n-1$ features. That is, each sample is a pair $(\mathbf{x}, y)$ containing a class label $y \in \{-1,+1\}$ and a vector of features $\mathbf{x} \in \mathbb{R}^{n-1}$. Let $\mathcal{Z}$ be the probability distribution of $(\mathbf{x}, y)$. That is, $(\mathbf{x}, y) \sim \mathcal{Z}$.

A linear classifier $f$ with weight vector $\mathbf{w} \in \mathbb{R}^n$ has the following decision function:

$$f(\mathbf{x}|\mathbf{w}) = \mathrm{sgn}(\mathbf{w}^{\mathrm{T}} \left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right]) \tag{2}$$

The linear classifier $f$ makes a mistake whenever the sign of $y$ does not match the sign of $f(\mathbf{x}|\mathbf{w})$, or equivalently, whenever:

$$y\mathbf{w}^{\mathrm{T}} \left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right] \leq 0 \tag{3}$$

In this paper, we analyze the probability of the above expression with respect to the data distribution. That is, we analyze the expected risk of a linear classifier.

Without loss of generality, let $\mathbf{z} = y\left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right]$. Clearly, eq.(3) holds if and only if $\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0$. We assume that $\mathbf{z}$ has mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$. That is:

$$\boldsymbol{\mu} \equiv \mathbb{E}[\mathbf{z}] \quad \text{and} \quad \boldsymbol{\Sigma} \equiv \mathbb{E}[(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}] \tag{4}$$

With some abuse of notation, we write $\mathbf{z} \sim \mathcal{Z} \equiv \mathcal{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in order to denote that the distribution $\mathcal{Z}$ of $\mathbf{z}$ has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Additionally, we define $\Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ to be the family of all distributions with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Thus, $\mathcal{Z} \in \Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is an equivalent way to denote that the distribution $\mathcal{Z}$ has mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Next, we introduce the following Fisher function of the weight vector $\mathbf{w}$, given the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the data distribution:

$$\mathcal{F}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \equiv \frac{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu})^2}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}} \tag{5}$$

The above expression is indeed the objective function of the Fisher linear discriminant, and it appears in the different scenarios under our analysis (cf. Theorems 1 and 4).

## 3 Bounds for the Expected Risk

The following "minimax" result was found by [16] and rediscovered by [3]. Proofs in both papers are more complex than the ones we provide here, and their main result is stated as follows. Let $\Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the family of all distributions with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. For a fixed weight vector $\mathbf{w}$ such that $\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$, we have:

$$\sup_{\mathcal{Z} \in \Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \mathbb{P}_{\mathbf{z}\sim\mathcal{Z}}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] = \frac{1}{1+\mathcal{F}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma})} \tag{6}$$

(See Appendix A for the relationship between this specific expression and the results in [3, 16].)

The "minimax" expression in eq.(6) motivated the development of "minimax probability machines" [13, 14], which only need access to the means and covariances for training. The *learning algorithm* in [13, 14] uses this bound for each class separately. That is, [13, 14] use the bound in eq.(6) with $\mathbf{x}$ (separately for class $y=-1$ and for class $y=+1$) instead of with $\mathbf{z}$, as in our setting. Note that [12, 14] also propose a *learning algorithm* called "single-class minimax probability machine", mainly motivated for the quantile estimation problem. For our *consistency analysis*, the use of a single bound with $\mathbf{z}$, allows us to provide closed form expressions of the upper bound of the expected risk.

Note that the "minimax" bound in eq.(6) provides an upper bound for every distribution $\mathcal{Z}$ and it shows that the bound is tight. In Theorems 1 and 2, we reproduce the same *polynomial-tail bounds* by following a simpler approach compared to semidefinite optimization [3] and multivariate constructions [16]. Our bound and tightness analysis relies on the analysis on a related univariate problem. In Theorems 4 and 5, we use a similar approach as we followed for the general random variables, and derive novel *exponential-tail bounds* for sub-Gaussian variates.

### 3.1 General Random Variables

First, we provide a bound of the expected risk of a *fixed* linear classifier for general random variables.

**Theorem 1.** *Let $\mathbf{z} = y\left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right]$ be a random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The expected risk of a linear classifier with weight vector $\mathbf{w}$ such that $\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$, is upper-bounded as follows:*

$$\mathbb{P}_{\mathbf{z}\sim\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] \leq \frac{1}{1+\mathcal{F}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{7}$$

*Proof.* Define the random variable $s = \mathbf{w}^{\mathrm{T}}\mathbf{z}$. It is easy to verify that $\mu_s \equiv \mathbb{E}[s] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$ and $\sigma_s^2 \equiv \mathbb{E}[(s-\mu_s)^2] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}$. By the one-sided Chebyshev's inequality for $\varepsilon = \mu_s > 0$, we have:

$$\begin{aligned}\mathbb{P}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] &= \mathbb{P}[s \leq 0]\\ &= \mathbb{P}[\mu_s - s \geq \varepsilon]\\ &\leq \tfrac{1}{1+(\varepsilon/\sigma_s)^2}\\ &= \tfrac{1}{1+(\mu_s/\sigma_s)^2}\end{aligned} \tag{8}$$

By replacing $\mu_s$ and $\sigma_s$, we prove our claim. □

Next, we show that the above bound is tight. That is, there is a data distribution and weight vector for which equality holds.

**Theorem 2.** *The upper bound provided in Theorem 1 is tight. That is, given an arbitrary mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, there is a distribution $\mathcal{Z}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and a weight vector $\mathbf{w}$ for which the bound holds with equality. More formally:*

$$\begin{pmatrix}\exists\mathcal{Z}\in\Omega(\boldsymbol{\mu},\boldsymbol{\Sigma})\\ \exists\mathbf{w}\in\mathbb{R}^n\end{pmatrix}\ \mathbb{P}_{\mathbf{z}\sim\mathcal{Z}}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] = \frac{1}{1+\mathcal{F}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{9}$$

*where $\Omega(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the family of all distributions with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.*

*Proof.* We show that given some arbitrary mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we can construct a distribution $\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ and a weight vector $\mathbf{w}$ such that the provided bound holds with equality. We prove this by providing a specific univariate "three-points" distribution which we later use for constructing a multivariate "three-planes" distribution.

First, we focus on the bound for a single scalar variable $s$ in eq.(8). Our goal is to construct a "three-points" distribution where $\mathbb{P}[\mu_s - s \geq \varepsilon] = \frac{1}{1+(\varepsilon/\sigma_s)^2}$. Thus, we want to show that the one-sided Chebyshev's inequality is tight. In fact, equality holds for the following distribution. Let $q(\varepsilon) = \frac{1}{4}\varepsilon(\varepsilon+\sqrt{\varepsilon^2+8})$ for some constant $\varepsilon \in (0, \sqrt{1/3})$ and $\beta > 0$, let $s$ be distributed as follows:

$$s = \begin{cases}\beta - 1, & \text{with probability } q(\varepsilon)\\ \beta, & \text{with probability } 1-2q(\varepsilon)\\ \beta+1, & \text{with probability } q(\varepsilon)\end{cases} \tag{10}$$

Note that $\mu_s \equiv \mathbb{E}[s] = \beta > 0$ and $\sigma_s^2 \equiv \mathbb{E}[(s-\mu_s)^2] = 2q(\varepsilon)$.

Second, we construct a canonical "three-planes" distribution. Consider a distribution $\mathcal{V}$ of random vectors $\mathbf{v}$, where $v_1 = (s+\beta)/\sqrt{2q(\varepsilon)}$ and $(v_2,\dots,v_n)$ follows a distribution with mean $\mathbf{0}$ and covariance $\mathbf{I}$. Since $v_1$ has zero mean and unit variance, as well as it is independent of $(v_2,\dots,v_n)$, we have $\mathbb{E}[\mathbf{v}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{v}\mathbf{v}^{\mathrm{T}}] = \mathbf{I}$. Let $\boldsymbol{\alpha} = (\sqrt{2q(\varepsilon)},0,\dots,0)$, we have $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{v}+\beta = s$ and therefore $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{v}+\beta$ is distributed as in eq.(10).

Finally, we construct a general "three-planes" distribution. For a given mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, we construct a random variable $\mathbf{z}$ with the required mean and covariance, from $\mathbf{v}$ as follows. Define the random variable $\mathbf{z} = \boldsymbol{\Sigma}^{1/2}\mathbf{v}+\boldsymbol{\mu}$. It is easy to verify that $\mathbb{E}[\mathbf{v}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{v}\mathbf{v}^{\mathrm{T}}] = \mathbf{I}$ if and only if $\mathbb{E}[\mathbf{z}] = \boldsymbol{\mu}$ and $\mathbb{E}[(\mathbf{z}-\boldsymbol{\mu})(\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}] = \boldsymbol{\Sigma}$. Note that $\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{v}+\beta = \mathbf{w}^{\mathrm{T}}\mathbf{z}$ for $\mathbf{w} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\alpha}$ and $\beta = \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu} > 0$. For $\varepsilon = \mu_s > 0$, we have:

$$\begin{aligned}\mathbb{P}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] &= \mathbb{P}[\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{v}+\beta \leq 0]\\ &= \mathbb{P}[s \leq 0]\\ &= \mathbb{P}[\mu_s - s \geq \varepsilon]\\ &= \tfrac{1}{1+(\varepsilon/\sigma_s)^2}\end{aligned}$$

and we prove our claim. □

### 3.2 Sub-Gaussian Random Variables

In this paper, we make use of the following definition of sub-Gaussianity of vectors by [7, 9].

**Definition 3.** *A random vector $\mathbf{z} \sim \mathcal{Z}$ is sub-Gaussian if for all $\mathbf{w} \in \mathbb{R}^n$, the random variable*

*$s = \mathbf{w}^{\mathrm{T}}\mathbf{z}$ with mean $\mu_s = \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[\mathbf{w}^{\mathrm{T}}\mathbf{z}]$ and variance $\sigma_s^2 = \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[(\mathbf{w}^{\mathrm{T}}\mathbf{z} - \mu_s)^2]$ is sub-Gaussian with parameter $\sigma_s$. That is:*

$$(\forall \mathbf{w} \in \mathbb{R}^n)\ \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[e^{\mathbf{w}^{\mathrm{T}}\mathbf{z}-\mu_s}] \le e^{\frac{1}{2}\sigma_s^2} \tag{11}$$

First, we provide a bound of the expected risk of a *fixed* linear classifier for sub-Gaussian random variables.

**Theorem 4.** *Let $\mathbf{z} = y\left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right]$ be a random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Assume $\mathbf{z}$ is a sub-Gaussian vector. The expected risk of a linear classifier with weight vector $\mathbf{w}$ such that $\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$, is upper-bounded as follows:*

$$\mathbb{P}_{\mathbf{z}\sim\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \le 0] \le e^{-\frac{1}{2}\mathcal{F}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{12}$$

*Proof.* Define the random variable $s = \mathbf{w}^{\mathrm{T}}\mathbf{z}$. It is easy to verify that $\mu_s \equiv \mathbb{E}[s] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$ and $\sigma_s^2 \equiv \mathbb{E}[(s-\mu_s)^2] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}$. Furthermore, by Definition 3, the random variable $s$ is sub-Gaussian with parameter $\sigma_s$. By the one-sided Chernoff bound for sub-Gaussian variables and $\varepsilon = \mu_s > 0$, we have:

$$\begin{aligned}\mathbb{P}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \le 0] &= \mathbb{P}[s \le 0]\\ &= \mathbb{P}[\mu_s - s \ge \varepsilon]\\ &\le e^{-\frac{1}{2}(\varepsilon/\sigma_s)^2}\\ &= e^{-\frac{1}{2}(\mu_s/\sigma_s)^2}\end{aligned} \tag{13}$$

By replacing $\mu_s$ and $\sigma_s$, we prove our claim. □

Next, we show that the above bound is tight. That is, there is a data distribution and weight vector for which equality holds.

**Theorem 5.** *The upper bound provided in Theorem 4 is tight. That is, given an arbitrary mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, there is a distribution $\mathcal{Z}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$, and a weight vector $\mathbf{w}$ for which the bound holds with equality. More formally:*

$$\begin{pmatrix}\exists \mathcal{Z} \in \Omega(\boldsymbol{\mu},\boldsymbol{\Sigma})\\ \exists \mathbf{w} \in \mathbb{R}^n\end{pmatrix}\ \mathbb{P}_{\mathbf{z}\sim\mathcal{Z}}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \le 0] = e^{-\frac{1}{2}\mathcal{F}(\mathbf{w}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{14}$$

*where $\Omega(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is the family of all distributions with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.*

*Proof.* A minor change to the "three-points" argument in Theorem 2 is needed. That is, we focus on the bound for a single scalar variable $s$ in eq.(13). Let $\mathcal{W}(a)$ be the Lambert function. That is, $\mathcal{W}(a)$ is the solution $t$ of the equation $a = te^t$. Our goal is to construct a "three-points" distribution where $\mathbb{P}[\mu_s - s \ge \varepsilon] = e^{-\frac{1}{2}(\varepsilon/\sigma_s)^2}$. Thus, we want to show that the one-sided Chernoff bound is tight. In fact, equality holds for the following distribution. Let $q(\varepsilon) = e^{\mathcal{W}(-\varepsilon^2/4)}$ for some constant $\varepsilon \in (0, \sqrt{2/e})$ and let $s$ be distributed as in eq.(10). Recall that bounded variables, such as the specifically constructed $s$, are sub-Gaussian. The rest of the proof follows as in Theorem 2 with the additional assumption that $\mathbf{v}$ is a sub-Gaussian vector. □

# 4 PAC-Bayes Finite-Sample Bounds

In this section, we provide finite-sample guarantees for our previously derived asymptotic bounds. Our "minimax" generalization bounds are dimensionality-independent and $\mathcal{O}(\sqrt{1/m})$ for $m$ samples, for both general and sub-Gaussian random variables.

First, we provide a brief introduction to the PAC-Bayes framework in the context of linear classifiers. Let $\mathcal{H}$ be a set of linear classifiers. That is, $\mathcal{H}$ is a set of weight vectors $\mathbf{w}$. After observing a training set, the task of the learner is to choose a posterior distribution of weight vectors $\mathcal{Q}$ of support $\mathcal{H}$, such that the Bayes classifier has the smallest possible risk. The Bayes linear classifier $f$ has the following decision function:

$$f(\mathbf{x}|\mathcal{Q}) = \mathrm{sgn}\left(\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right]]\right) \tag{15}$$

The output of the (deterministic) Bayes classifier is closely related to the output of the (stochastic) Gibbs classifier. The Gibbs linear classifier chooses randomly a (deterministic) classifier $\mathbf{w}$ according to $\mathcal{Q}$ in order to classify $\mathbf{x}$. The expected risk of the Gibbs linear classifier is thus given by the probability of eq.(3), with respect to the data distribution and the posterior distribution $\mathcal{Q}$.

PAC-Bayes guarantees are given with respect to a prior distribution of weight vectors $\mathcal{P}$, also of support $\mathcal{H}$, and mostly focus on the analysis of the Gibbs classifier. As noted in [8], the expected risk of the Bayes classifier is at most twice the expected risk of the Gibbs classifier. Thus, any upper bound of the latter provides an upper bound of the former. The factor of 2 can sometimes be reduced to $1+\varepsilon$, as shown in [15].

In our analysis, we chose a specific $n$-dimensional ellipsoid as the support:

$$\mathcal{H} = \{\mathbf{w} \mid \mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w} = 1\} \tag{16}$$

The latter is for convenience. Since the Fisher function uses only linear and quadratic terms, any distribution of support $\mathbb{R}^n$ can be reparametrized with respect to $\mathcal{H}$ and produce the same results. For instance, the Fisher function is scale-independent with respect to $\mathbf{w}$. Intuitively speaking, the reparametrization can be performed by integrating the mass of the distribution of support $\mathbb{R}^n$ over every direction independently.

Indeed, the distributions $\mathcal{P}$ and $\mathcal{Q}$ could also be described with respect to angles. Fortunately, we do not need such a reparamerization for our analysis.

Next, we introduce the following Gibbs-Fisher function of weight vectors $\mathbf{w}$ sampled from a distribution $\mathcal{Q}$, given the mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ of the data distribution:

$$\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \equiv \frac{(\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2}{\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w}]-(\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2} \tag{17}$$

The above expression appears in the bound of the expected risk of the Gibbs linear classifier, in the different scenarios under our analysis (cf. Theorems 6 and 11).

Our exponential-tail bounds for sub-Gaussian variates depend on the Kullback-Leibler divergence, while our polynomial-tail bounds for general random variables depend on the Chi-squared divergence. Given two distributions $\mathcal{P}$ and $\mathcal{Q}$, the Kullback-Leibler and Chi-squared divergences are defined as $\mathcal{KL}(\mathcal{Q}||\mathcal{P}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\log\frac{q(\mathbf{w})}{p(\mathbf{w})}]$ and $\chi^2(\mathcal{Q}||\mathcal{P}) = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\frac{q(\mathbf{w})}{p(\mathbf{w})}] - 1$, respectively.

Our proof strategy is to first show concentration of the *projected* mean and variance, and then use those results in order to show concentration of the Gibbs-Fisher function.

### 4.1 General Random Variables

First, we provide a bound of the expected risk of the Gibbs linear classifier for general random variables.

**Theorem 6.** *Let* $\mathbf{z} = y\left[\begin{smallmatrix}\mathbf{x}\\1\end{smallmatrix}\right]$ *be a random variable with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$. *Let* $\mathcal{Q}$ *be the probability distribution of the random weight vector* $\mathbf{w}$. *The expected risk of a linear classifier drawn from* $\mathcal{Q}$ *such that* $\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}] > 0$, *is upper-bounded as follows:*

$$\mathbb{P}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \le 0] \le \frac{1}{1+\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{18}$$

*Proof.* Define the random variable $s = \mathbf{w}^{\mathrm{T}}\mathbf{z}$. It is easy to verify that $\mu_s \equiv \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[s] = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}] > 0$ and $\sigma_s^2 \equiv \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[(s-\mu_s)^2] = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w}] - (\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2$. The rest of the proof follows as in Theorem 1. □

Next, we show PAC-Bayes concentration of the *projected* mean for general random variables.

**Lemma 7.** *Let* $\mathbf{z}^{(1)},\dots,\mathbf{z}^{(m)}$ *be* $m$ *samples independently drawn from* $\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, *an arbitrary distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$. *Let* $\widehat{\boldsymbol{\mu}}$ *be the empirical mean computed from those samples. For any prior distribution* $\mathcal{P}$ *of support* $\mathcal{H}$ *as in eq.(16), with probability at least* $1-\delta$:

$$(\forall\mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})]\right| \le \sqrt{\frac{1}{m\delta}(\chi^2(\mathcal{Q}||\mathcal{P})+1)} \tag{19}$$

*Proof.* Note that the random variable $\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2]$ is non-negative. By Markov's inequality, with probability at least $1-\delta$:

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2]\\ &\qquad\le \frac{1}{\delta}\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2]\end{aligned} \tag{20}$$

Define the random variable $s = \mathbf{w}^{\mathrm{T}}(\mathbf{z}-\boldsymbol{\mu})$. It is easy to verify that $\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[s] = 0$ and $\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[s^2] = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w} \le \mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w} = 1$ in the support $\mathcal{H}$ as in eq.(16).

Note that $\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}) = \frac{1}{m}\sum_{i=1}^m s^{(i)}$. Furthermore, $s^{(1)},\dots,s^{(m)}$ are independent since $\mathbf{z}^{(1)},\dots,\mathbf{z}^{(m)}$ are independent. Thus, we can upper-bound the expected value in the right-hand side of eq.(20) as follows:

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2]\\ &= \mathbb{E}_{\mathbf{w}\sim\mathcal{P}}\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\left[\left(\tfrac{1}{m}\textstyle\sum_{i=1}^m s^{(i)}\right)^2\right]\\ &= \mathbb{E}_{\mathbf{w}\sim\mathcal{P}}\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\left[\tfrac{1}{m^2}\left(\textstyle\sum_{i=1}^m {s^{(i)}}^2 + \sum_{i\ne j} s^{(i)}s^{(j)}\right)\right]\\ &= \mathbb{E}_{\mathbf{w}\sim\mathcal{P}}\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}\left[s^2/m\right]\\ &\le 1/m\end{aligned}$$

By putting everything together, we have:

$$\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2] \le 1/(m\delta) \tag{21}$$

By Jensen's and Cauchy-Schwarz inequalities, we have:

$$\begin{aligned}&(\forall\mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})]\right|\\ &\quad\le \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\left|\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right|]\\ &\quad= \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\left(\sqrt{\tfrac{p(\mathbf{w})}{q(\mathbf{w})}}\left|\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right|\right)\sqrt{\tfrac{q(\mathbf{w})}{p(\mathbf{w})}}\right]\\ &\quad\le \sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\tfrac{p(\mathbf{w})}{q(\mathbf{w})}(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2\right]\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\tfrac{q(\mathbf{w})}{p(\mathbf{w})}\right]}\\ &\quad= \sqrt{\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu}))^2]\ (\chi^2(\mathcal{Q}||\mathcal{P})+1)}\end{aligned}$$

By using our bound in eq.(21), we prove our claim. □

In what follows, we show PAC-Bayes concentration of the *projected* variance for general random variables.

**Lemma 8.** *Let* $\mathbf{z}^{(1)},\dots,\mathbf{z}^{(m)}$ *be* $m$ *samples independently drawn from* $\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, *an arbitrary distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$. *Let* $\widehat{\boldsymbol{\mu}}$ *and* $\widehat{\boldsymbol{\Sigma}}$ *be the empirical mean and covariance computed from those samples. Assume* $\mathbf{z}$ *has bounded fourth order moment, that is* $\mathbb{E}[(\mathbf{z}^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})^{-1}\mathbf{z})^2] \le K$. *For any prior* *distribution $\mathcal{P}$ of support $\mathcal{H}$ as in eq.*(16)*, with probability at least $1-\delta$:*

$$(\forall \mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1]\right| \le \sqrt{\frac{K-1}{m\delta}(\chi^2(\mathcal{Q}||\mathcal{P})+1)} \tag{22}$$

*Proof.* Note that the random variable $\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)^2]$ is non-negative. By Markov's inequality, with probability at least $1-\delta$:

$$\begin{aligned}&\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)^2]\\ &\le \frac{1}{\delta}\mathbb{E}_{\mathbf{z}^{(1)}\ldots\mathbf{z}^{(m)}\sim\mathcal{Z}}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)^2]\end{aligned} \tag{23}$$

Define the random variable $s=(\mathbf{w}^{\mathrm{T}}\mathbf{z})^2-1$. Note that in the support $\mathcal{H}$ as in eq.(16), we have $\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[s]=\mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w}-1=0$. Additionally, let $\mathbf{S}=\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$ and by our assumption of bounded fourth order moment, we have:

$$\begin{aligned}\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[s^2]&=\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[(\mathbf{w}^{\mathrm{T}}\mathbf{z})^4]-1\\ &=\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[((\mathbf{S}^{1/2}\mathbf{w})^{\mathrm{T}}\mathbf{S}^{-1/2}\mathbf{z})^4]\\ &\le \|\mathbf{S}^{1/2}\mathbf{w}\|_2^4\ \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[\|\mathbf{S}^{-1/2}\mathbf{z}\|_2^4]-1\\ &=(\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w})^2\ \mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[(\mathbf{z}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{z})^2]-1\\ &\le K-1\end{aligned}$$

Note that $\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1=\frac{1}{m}\sum_{i=1}^m s^{(i)}$. Furthermore, $s^{(1)},\ldots,s^{(m)}$ are independent since $\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)}$ are independent. Thus, we can upper-bound the expected value in the right-hand side of eq.(23). The rest of the proof follows similarly as in Lemma 7. □

Finally, we show PAC-Bayes concentration of the Gibbs-Fisher function for general random variables. Our "minimax" generalization bound is dimensionality-independent and $\mathcal{O}(\sqrt{1/m})$ for $m$ samples.

**Theorem 9.** *Let $\mathbf{z}^{(1)},\ldots,\mathbf{z}^{(m)}$ be $m$ samples independently drawn from $\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, an arbitrary distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Let $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the empirical mean and covariance computed from those samples. Assume $\mathbf{z}$ has bounded fourth order moment, that is $\mathbb{E}[(\mathbf{z}^{\mathrm{T}}(\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})^{-1}\mathbf{z})^2]\le K$. For any prior distribution $\mathcal{P}$ of support $\mathcal{H}$ as in eq.*(16)*, with probability at least $1-\delta$:*

$$\begin{aligned}&(\forall \mathcal{Q})\ \left|\frac{1}{1+\mathcal{F}(\mathcal{Q}|\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})}-\frac{1}{1+\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\right|\\ &\le \sqrt{\frac{18\max{(1,K-1)}}{m\delta}(\chi^2(\mathcal{Q}||\mathcal{P})+1)}+\mathcal{O}\left(\frac{1}{m}\right)\end{aligned} \tag{24}$$

*Proof.* In order to obtain concentration of the *projected* mean and variance simultaneously, we apply the union bound to Lemmas 7 and 8. That is, with probability at least $1-\delta$, let $\varepsilon=\sqrt{\frac{2\max{(1,K-1)}}{m\delta}(\chi^2(\mathcal{Q}||\mathcal{P})+1)}$, we have:

$$\begin{aligned}&(\forall \mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})]\right|\le\varepsilon\\ &(\forall \mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1]\right|\le\varepsilon\end{aligned} \tag{25}$$

With some algebra, we can prove that for any $\mathcal{Q}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{S}=\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$, we have:

$$\frac{1}{1+\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma})}=1-\frac{(\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2}{\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}]} \tag{26}$$

Let $\widehat{\mathbf{S}}=\widehat{\boldsymbol{\Sigma}}+\widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}}$, $\alpha=\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}]$, $\widehat{\alpha}=\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}]$ and $\widehat{\beta}=\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\widehat{\mathbf{S}}\mathbf{w}]$. The concentration results in eq.(25) are equivalent to $|\widehat{\alpha}-\alpha|\le\varepsilon$ and $|\widehat{\beta}-1|\le\varepsilon$. Additionally by eq.(26), the left-hand side of eq.(24) is equivalent to $|\widehat{\alpha}^2/\widehat{\beta}-\alpha^2|$.

For any $\mathcal{Q}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{S}=\boldsymbol{\Sigma}+\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$, by positive semidefiniteness of $\boldsymbol{\Sigma}$, we have:

$$\begin{aligned}(\forall \mathbf{w})\ 0\le\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}&\Rightarrow 0\le\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}]\\ &\Rightarrow 0\le\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\mathbf{S}-\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w}]\\ &\Rightarrow \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu})^2]\le\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}]\\ &\Rightarrow \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu})^2]/\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}]\le 1\end{aligned}$$

Note that $0\le(\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2\le\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu})^2]$ and therefore $\widehat{\alpha}^2/\widehat{\beta}\in[0,1]$ and $\alpha^2\in[0,1]$. Since $|\widehat{\alpha}-\alpha|\le\varepsilon$ and $\alpha\le 1$, we have:

$$\begin{aligned}\widehat{\alpha}^2-\alpha^2&=2\alpha(\widehat{\alpha}-\alpha)+(\widehat{\alpha}-\alpha)^2\\ &\le 2\varepsilon+\varepsilon^2\end{aligned}$$

Similarly:

$$\begin{aligned}\alpha^2-\widehat{\alpha}^2&=2\widehat{\alpha}(\alpha-\widehat{\alpha})+(\alpha-\widehat{\alpha})^2\\ &\le 2(\alpha+\varepsilon)\varepsilon+\varepsilon^2\\ &\le 2(1+\varepsilon)\varepsilon+\varepsilon^2\\ &=2\varepsilon+3\varepsilon^2\end{aligned}$$

Therefore, $|\widehat{\alpha}^2-\alpha^2|\le 2\varepsilon+3\varepsilon^2$. Finally, since $|\widehat{\alpha}-\alpha|\le\varepsilon$, $|\widehat{\beta}-1|\le\varepsilon$ and $\widehat{\alpha}^2/\widehat{\beta}\le 1$, we have:

$$\begin{aligned}|\widehat{\alpha}^2/\widehat{\beta}-\alpha^2|&\le|\widehat{\alpha}^2/\widehat{\beta}-\widehat{\alpha}^2|+|\widehat{\alpha}^2-\alpha^2|\\ &=\widehat{\alpha}^2/\widehat{\beta}|1-\widehat{\beta}|+|\widehat{\alpha}^2-\alpha^2|\\ &\le\varepsilon+2\varepsilon+3\varepsilon^2\\ &=3\varepsilon+\mathcal{O}(1/m)\end{aligned}$$

and we prove our claim. □

### 4.2 Sub-Gaussian Random Variables

In this paper, we introduce the following definition of sub-Gaussianity of vectors.

**Definition 10.** *Let $\mathcal{H}$ be a bounded set. A random vector $\mathbf{z} \sim \mathcal{Z}$ is $\mathcal{H}$-*sub-Gaussian *if for all distributions $\mathcal{Q}$ of support $\mathcal{H}$ and $\mathbf{w} \sim \mathcal{Q}$, the random variable $s = \mathbf{w}^{\mathrm{T}}\mathbf{z}$ with mean $\mu_s = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[\mathbf{w}^{\mathrm{T}}\mathbf{z}]$ and variance $\sigma_s^2 = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[(\mathbf{w}^{\mathrm{T}}\mathbf{z} - \mu_s)^2]$ is sub-Gaussian with parameter $\sigma_s$. That is:*

$$(\forall \mathcal{Q})\ \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[e^{\mathbf{w}^{\mathrm{T}}\mathbf{z}-\mu_s}] \leq e^{\frac{1}{2}\sigma_s^2} \tag{27}$$

First, we provide a bound of the expected risk of the Gibbs linear classifier for sub-Gaussian random variables.

**Theorem 11.** *Let $\mathbf{z} = y\left[\begin{smallmatrix}\mathbf{x}\\ 1\end{smallmatrix}\right]$ be a random variable with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Let $\mathcal{Q}$ be the probability distribution of the random weight vector $\mathbf{w}$ of support $\mathcal{H}$. Assume $\mathbf{z}$ is an $\mathcal{H}$-sub-Gaussian vector. The expected risk of a linear classifier drawn from $\mathcal{Q}$ such that $\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}] > 0$, is upper-bounded as follows:*

$$\mathbb{P}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] \leq e^{-\frac{1}{2}\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma})} \tag{28}$$

*Proof.* Define the random variable $s = \mathbf{w}^{\mathrm{T}}\mathbf{z}$. It is easy to verify that $\mu_s \equiv \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[s] = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}] > 0$ and $\sigma_s^2 \equiv \mathbb{E}_{\mathbf{w}\sim\mathcal{Q},\mathbf{z}\sim\mathcal{Z}}[(s-\mu_s)^2] = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w}] - (\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2$. Furthermore, by Definition 10, the random variable $s$ is sub-Gaussian with parameter $\sigma_s$. The rest of the proof follows as in Theorem 4. □

Next, we show PAC-Bayes concentration of the *projected* mean for sub-Gaussian random variables. In the following lemma, while we concentrate on upper-bounding $\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})]$ for all $\mathcal{Q}$, a similar argument also bounds $\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}})]$.

**Lemma 12.** *Let $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ be $m$ samples independently drawn from $\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, an arbitrary distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Let $\widehat{\boldsymbol{\mu}}$ be the empirical mean computed from those samples. Assume $\mathbf{z}$ is a sub-Gaussian vector. For any prior distribution $\mathcal{P}$ of support $\mathcal{H}$ as in eq.(16), with probability at least $1-\delta$:*

$$(\forall \mathcal{Q})\ \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})] \leq \sqrt{\frac{1}{m}}\left(\mathcal{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{e^{1/2}}{\delta}\right) \tag{29}$$

*Proof.* Let $t \in \mathbb{R}$ be a constant. Note that the random variable $\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}]$ is non-negative. By Markov's inequality, with probability at least $1-\delta$:

$$\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}] \leq \frac{1}{\delta}\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}] \tag{30}$$

Define the random variable $s = \mathbf{w}^{\mathrm{T}}(\mathbf{z}-\boldsymbol{\mu})$. It is easy to verify that $\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[s] = 0$. As we argued in Theorem 4, the variable $s$ is sub-Gaussian with parameter $\sigma_s$ where $\sigma_s^2 = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}$. Furthermore, $\sigma_s^2 \leq \mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w} = 1$ in the support $\mathcal{H}$ as in eq.(16).

Note that $\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu}) = \frac{1}{m}\sum_{i=1}^{m} s^{(i)}$. Furthermore, $s^{(1)}, \dots, s^{(m)}$ are independent since $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ are independent. Thus, we can upper-bound the expected value in the right-hand side of eq.(30) as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}] \\
&\qquad = \mathbb{E}_{\mathbf{w}\sim\mathcal{P}}\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}[e^{\frac{t}{m}\sum_{i=1}^{m} s^{(i)}}] \\
&\qquad = \mathbb{E}_{\mathbf{w}\sim\mathcal{P}}\prod_{i=1}^{m}\mathbb{E}_{\mathbf{z}\sim\mathcal{Z}}[e^{\frac{t}{m}s}] \\
&\qquad \leq \mathbb{E}_{\mathbf{w}\sim\mathcal{P}}\prod_{i=1}^{m} e^{\frac{t^2}{2m^2}} \\
&\qquad = e^{\frac{t^2}{2m}}
\end{aligned}$$

By taking the logarithm on each side of eq.(30) and since $\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[f(\mathbf{w})] = \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\frac{p(\mathbf{w})}{q(\mathbf{w})}f(\mathbf{w})]$ for every distribution $\mathcal{Q}$ and function $f : \mathbb{R}^n \to R$, we have:

$$\begin{aligned}
&(\forall \mathcal{Q})\ \log\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\tfrac{p(\mathbf{w})}{q(\mathbf{w})}e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}\right] \\
&\qquad \leq \log\left(\tfrac{1}{\delta}\mathbb{E}_{\mathbf{z}^{(1)}\dots\mathbf{z}^{(m)}\sim\mathcal{Z}}\mathbb{E}_{\mathbf{w}\sim\mathcal{P}}[e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}]\right)
\end{aligned}$$

By using Jensen's inequality, we can lower-bound the left-hand side of the above expression:

$$\begin{aligned}
&(\forall \mathcal{Q})\ \log\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\tfrac{p(\mathbf{w})}{q(\mathbf{w})}e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}\right] \\
&\qquad \geq \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}\left[\log\left(\tfrac{p(\mathbf{w})}{q(\mathbf{w})}e^{t\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})}\right)\right] \\
&\qquad = -\mathcal{KL}(\mathcal{Q}||\mathcal{P}) + t\ \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})]
\end{aligned}$$

By putting everything together, we have:

$$(\forall \mathcal{Q})\ \mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}-\boldsymbol{\mu})] \leq \tfrac{1}{t}\left(\mathcal{KL}(\mathcal{Q}||\mathcal{P}) + \log\frac{e^{\frac{t^2}{2m}}}{\delta}\right)$$

By setting $t = \sqrt{m}$, we prove our claim. □

In what follows, we show PAC-Bayes concentration of the *projected* variance for sub-Gaussian random variables. In the following lemma, while we concentrate on upper-bounding $\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w} - 1]$ for all $\mathcal{Q}$, a similar argument also bounds $\mathbb{E}_{\mathbf{w}\sim\mathcal{Q}}[1 - \mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}]$.

**Lemma 13.** *Let $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}$ be $m \geq 16$ samples independently drawn from $\mathcal{Z}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, an arbitrary distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Let $\widehat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ be the empirical mean and covariance computed from those samples. Assume $\mathbf{z}$ is a sub-Gaussian vector. For any prior distribution $\mathcal{P}$ of support $\mathcal{H}$ as in eq.(16),* *with probability at least* $1 - \delta$*:*

$$(\forall \mathcal{Q})\ \mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w} - 1] \leq \sqrt{\frac{1}{m}\left(\mathcal{KL}(\mathcal{Q}||\mathcal{P}) + \log \frac{e^{16}}{\delta}\right)} \tag{31}$$

*Proof.* Let $t \in \mathbb{R}$ be a constant. Note that the random variable $\mathbb{E}_{\mathbf{w} \sim \mathcal{P}}[e^{t(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)}]$ is non-negative. By Markov's inequality, with probability at least $1 - \delta$:

$$\mathbb{E}_{\mathbf{w} \sim \mathcal{P}}[e^{t(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)}] \leq \frac{1}{\delta}\mathbb{E}_{\mathbf{z}^{(1)} \ldots \mathbf{z}^{(m)} \sim \mathcal{Z}}\mathbb{E}_{\mathbf{w} \sim \mathcal{P}}[e^{t(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)}] \tag{32}$$

Define the random variable $s = (\mathbf{w}^{\mathrm{T}}\mathbf{z})^2 - 1$. As we argued in Theorem 4, the random variable $\mathbf{w}^{\mathrm{T}}\mathbf{z}$ is sub-Gaussian with parameter $\sigma$ where $\sigma^2 = \mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}$. Furthermore, $\sigma^2 \leq \mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w} = 1$ in the support $\mathcal{H}$ as in eq.(16). Additionally, in the support $\mathcal{H}$, we have $\mathbb{E}_{\mathbf{z} \sim \mathcal{Z}}[s] = \mathbf{w}^{\mathrm{T}}(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}})\mathbf{w} - 1 = 0$ and furthermore, $s$ is sub-exponential since it is the square of a sub-Gaussian variable. In what follows, we use a bound for the moment generating function of a sub-exponential variable, that resembles that of sub-Gaussian variables. (See Appendix B for details.)

Note that $\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w} - 1 = \frac{1}{m}\sum_{i=1}^m s^{(i)}$. Furthermore, $s^{(1)}, \ldots, s^{(m)}$ are independent since $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ are independent. Thus, we can upper-bound the expected value in the right-hand side of eq.(32) as follows:

$$\begin{aligned}
&\mathbb{E}_{\mathbf{z}^{(1)} \ldots \mathbf{z}^{(m)} \sim \mathcal{Z}}\mathbb{E}_{\mathbf{w} \sim \mathcal{P}}[e^{t(\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w}-1)}] \\
&\quad = \mathbb{E}_{\mathbf{w} \sim \mathcal{P}}\mathbb{E}_{\mathbf{z}^{(1)} \ldots \mathbf{z}^{(m)} \sim \mathcal{Z}}[e^{\frac{t}{m}\sum_{i=1}^m s^{(i)}}] \\
&\quad = \mathbb{E}_{\mathbf{w} \sim \mathcal{P}} \textstyle\prod_{i=1}^m \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}}[e^{\frac{t}{m}s}] \\
&\quad \leq \mathbb{E}_{\mathbf{w} \sim \mathcal{P}} \textstyle\prod_{i=1}^m e^{\frac{16t^2}{m^2}} \text{ for } |{}^t/_m| \leq {}^1/_4 \\
&\quad = e^{\frac{16t^2}{m}}
\end{aligned}$$

The rest of the proof follows similarly as in Lemma 12. Note that since we set $t = \sqrt{m}$, the condition $|{}^t/_m| \leq {}^1/_4$ for applying the bound for the moment generating function of the sub-exponential variable $s$ leads to the condition $m \geq 16$. □

Finally, we show PAC-Bayes concentration of the Gibbs-Fisher function for sub-Gaussian random variables. Our "minimax" generalization bound is dimensionality-independent and $\mathcal{O}(\sqrt{{}^1/_m})$ for $m$ samples.

**Theorem 14.** *Let* $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$ *be* $m \geq 16$ *samples independently drawn from* $\mathcal{Z}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*, an arbitrary distribution with mean* $\boldsymbol{\mu}$ *and covariance* $\boldsymbol{\Sigma}$*. Let* $\widehat{\boldsymbol{\mu}}$ *and* $\widehat{\boldsymbol{\Sigma}}$ *be the empirical mean and covariance computed from those samples. Assume* $\mathbf{z}$ *is a sub-Gaussian vector. For any prior distribution* $\mathcal{P}$ *of support* $\mathcal{H}$ *as in eq.*(16)*, with probability at least* $1 - \delta$*:*

$$(\forall \mathcal{Q})\ \left|e^{-\frac{1}{2}\mathcal{F}(\mathcal{Q}|\widehat{\boldsymbol{\mu}},\widehat{\boldsymbol{\Sigma}})} - e^{-\frac{1}{2}\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma})}\right| \leq \sqrt{\frac{36}{m}\left(\mathcal{KL}(\mathcal{Q}||\mathcal{P}) + \log \frac{4e^{16}}{\delta}\right)} + \mathcal{O}\left(\frac{1}{m}\right) \tag{33}$$

*Proof.* In order to obtain *two-sided* concentration of both, the *projected* mean and variance simultaneously, we apply the union bound to the *one-sided* results in Lemmas 12 and 13. That is, with probability at least $1 - \delta$, let $\varepsilon = \sqrt{\frac{1}{m}\left(\mathcal{KL}(\mathcal{Q}||\mathcal{P}) + \log \frac{4e^{16}}{\delta}\right)}$, we have:

$$\begin{aligned}
&(\forall \mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\mu}} - \boldsymbol{\mu})]\right| \leq \varepsilon \\
&(\forall \mathcal{Q})\ \left|\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}(\widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}})\mathbf{w} - 1]\right| \leq \varepsilon
\end{aligned} \tag{34}$$

With some algebra, we can prove that for any $\mathcal{Q}$, $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\mathbf{S} = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}$, we have:

$$e^{-\frac{1}{2}\mathcal{F}(\mathcal{Q}|\boldsymbol{\mu},\boldsymbol{\Sigma})} = \exp\left(-\frac{\frac{(\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2}{\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}]}}{2\left(1 - \frac{(\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}])^2}{\mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\mathbf{S}\mathbf{w}]}\right])}\right) \tag{35}$$

Let $\widehat{\mathbf{S}} = \widehat{\boldsymbol{\Sigma}} + \widehat{\boldsymbol{\mu}}\widehat{\boldsymbol{\mu}}^{\mathrm{T}}$, $\alpha = \mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}]$, $\widehat{\alpha} = \mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\widehat{\boldsymbol{\mu}}]$ and $\widehat{\beta} = \mathbb{E}_{\mathbf{w} \sim \mathcal{Q}}[\mathbf{w}^{\mathrm{T}}\widehat{\mathbf{S}}\mathbf{w}]$. The concentration results in eq.(34) are equivalent to $|\widehat{\alpha} - \alpha| \leq \varepsilon$ and $|\widehat{\beta} - 1| \leq \varepsilon$. Note that by eq.(35), the left-hand side of eq.(33) is equivalent to $\left|\exp\left(-\frac{\widehat{\alpha}^2/\widehat{\beta}}{2(1-\widehat{\alpha}^2/\widehat{\beta})}\right) - \exp\left(-\frac{\alpha^2}{2(1-\alpha^2)}\right)\right|$. Furthermore, by Lipschitz continuity:

$$\left|\exp\left(-\frac{\widehat{\alpha}^2/\widehat{\beta}}{2(1-\widehat{\alpha}^2/\widehat{\beta})}\right) - \exp\left(-\frac{\alpha^2}{2(1-\alpha^2)}\right)\right| \leq 2|\widehat{\alpha}^2/\widehat{\beta} - \alpha^2|$$

The rest of the proof follows as in Theorem 9 to show that $|\widehat{\alpha}^2/\widehat{\beta} - \alpha^2| \leq 3\varepsilon + \mathcal{O}(1/m)$. □

## 5 Concluding Remarks

There are several ways of extending this research. While in this paper we focused in the PAC-Bayes framework, one future goal is to provide guarantees for all weight vectors $\mathbf{w}$. In order to obtain bounds that are either independent or logarithmically-dependent on the dimension, it might be necessary to produce novel bounds for the Rademacher and/or Gaussian complexity of *not-everywhere-bounded* functions. Finally, since our bounds on the expected risk are worst-case (among all data distributions with a given mean and covariance), it would be interesting to analyze their applicability to scenarios where each training sample may come from a different distribution, as well as for transfer learning.

## References

[1] M. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. *COLT*, 2013.

[2] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*, 2002.

[3] D. Bertsimas, I. Popescu, and J. Sethuraman. Moment problems and semidefinite optimization. *Handbook of Semidefinite Optimization*, 2000.

[4] O. Bousquet, S. Boucheron, and G. Lugosi. Introduction to statistical learning theory. *Advanced Lectures on Machine Learning*, 2004.

[5] O. Bousquet and A. Elisseeff. Stability and generalization. *JMLR*, 2002.

[6] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. IMAGENET: A large-scale hierarchical image database. *CVPR*, 2009.

[7] R. Fukuda. Exponential integrability of subgaussian vectors. *Probability Theory and Related Fields*, 1990.

[8] P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. *ICML*, 2009.

[9] D. Hsu, S. Kakade, and T. Zhang. A tail inequality for quadratic forms of sub-Gaussian random vectors. *Electronic Communications in Probability*, 2012.

[10] C. Jin and L. Wang. Dimensionality dependent PAC-Bayes margin bound. *NIPS*, 2012.

[11] S. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. *NIPS*, 2008.

[12] G. Lanckriet, L. El Ghaoui, , and M. Jordan. Robust novelty detection with single-class MPM. *NIPS*, 2002.

[13] G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. Jordan. Minimax probability machine. *NIPS*, 2001.

[14] G. Lanckriet, L. El Ghaoui, C. Bhattacharyya, and M. Jordan. A robust minimax approach to classification. *JMLR*, 2002.

[15] J. Langford and J. Shawe-Taylor. PAC-Bayes & margins. *NIPS*, 2002.

[16] A. Marshall and I. Olkin. Multivariate Chebyshev inequalities. *The Annals of Mathematical Statistics*, 1960.

[17] D. McAllester. Simplified PAC-Bayesian margin bounds. *COLT*, 2003.

[18] S. Mukherjee, P. Niyogi, T. Poggio, and R. Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 2006.

[19] A. Ng. Feature selection, $\ell_1$ vs. $\ell_2$ regularization, and rotational invariance. *ICML*, 2004.

[20] S. Rakhlin, S. Mukherjee, and T. Poggio. Stability results in learning theory. *Analysis and Applications*, 2005.

[21] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *JMLR*, 2010.

[22] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: a large dataset for nonparametric object and scene recognition. *PAMI*, 2008.

[23] T. Zhang. Covering number bounds of certain regularized linear function classes. *JMLR*, 2002.

# A Proof of Expression in Equation (6)

In this section, we restate the results provided in [3, 16] in order to obtain eq.(6). We follow the well-known Lagrangian duality approach as in Appendix A in [14].

The following result is provided in [3, 16]. Let $\Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the family of all distributions with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. For a fixed weight vector $\mathbf{a}$ and constant $b$, we have:

$$\begin{aligned} \sup_{\mathcal{Z} \in \Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \quad & \mathbb{P}_{\mathbf{z} \sim \mathcal{Z}}[\mathbf{a}^{\mathrm{T}}\mathbf{z} \geq b] = \frac{1}{1+d^2} \\ \text{where} \quad & d^2 = \inf_{\mathbf{a}^{\mathrm{T}}\mathbf{z} \geq b} (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \end{aligned}$$

Let $\mathbf{a} = -\mathbf{w}$ and $b = 0$. We have:

$$\begin{aligned} \sup_{\mathcal{Z} \in \Omega(\boldsymbol{\mu}, \boldsymbol{\Sigma})} \quad & \mathbb{P}_{\mathbf{z} \sim \mathcal{Z}}[\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0] = \frac{1}{1+d^2} \\ \text{where} \quad & d^2 = \inf_{\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0} (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \end{aligned}$$

Note that if $\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} \leq 0$, then we can just take $\mathbf{z} = \boldsymbol{\mu}$ and obtain $d^2 = 0$, which is certainly the optimum because $d^2 \geq 0$ due to positive definiteness of $\boldsymbol{\Sigma}$. In what follows, we assume $\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$, as required in eq.(6).

We are interested in the value of $d^2$. That is, we seek for a closed-form solution of the *primal* problem:

$$\min_{\mathbf{w}^{\mathrm{T}}\mathbf{z} \leq 0} (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) \tag{36}$$

which has the following Lagrangian:

$$\mathcal{L}(\mathbf{z}, \lambda) = (\mathbf{z}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{z}-\boldsymbol{\mu}) + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{z}$$

By optimality arguments (i.e. $\partial\mathcal{L}/\partial\mathbf{z} = \mathbf{0}$), we have that $\mathcal{L}$ is minimized at $\mathbf{z}^* = -\frac{\lambda}{2}\boldsymbol{\Sigma}\mathbf{w} + \boldsymbol{\mu}$. Therefore, the Lagrange dual function is given by:

$$\begin{aligned} g(\lambda) &= \inf_{\mathbf{z}} \mathcal{L}(\mathbf{z}, \lambda) \\ &= \mathcal{L}(\mathbf{z}^*, \lambda) \\ &= -\frac{\lambda^2}{4}\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w} + \lambda\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} \end{aligned}$$

Consequently, the *dual* problem of eq.(36) is:

$$\max_{\lambda \geq 0} g(\lambda)$$

Again, by optimality arguments (i.e. $\partial g/\partial\lambda = 0$), we have that $g$ is maximized at $\lambda^* = 2\frac{\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu}}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}}$. Note that $\lambda^* \geq 0$ since $\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu} > 0$. Finally:

$$\begin{aligned} d^2 &= \max_{\lambda \geq 0} g(\lambda) \\ &= g(\lambda^*) \\ &= \frac{(\mathbf{w}^{\mathrm{T}}\boldsymbol{\mu})^2}{\mathbf{w}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{w}} \\ &\equiv \mathcal{F}(\mathbf{w}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \end{aligned}$$

# B Moment Generating Function of the Square of a Sub-Gaussian Variable

Let $s$ be a sub-Gaussian variable with parameter $\sigma_s$ and mean $\mu_s = \mathbb{E}[s]$. By sub-Gaussianity, we know that the moment generating function is bounded as follows:

$$(\forall t \in \mathbb{R}) \ \mathbb{E}[e^{t(s-\mu_s)}] \leq e^{\frac{1}{2}t^2\sigma_s^2}$$

Our goal is to find a similar bound for the moment generating function of the sub-exponential variable $v = s^2$. Let $\Gamma(r)$ be the Gamma function, the moments of the sub-Gaussian variable $s$ are bounded as follows:

$$(\forall r \geq 0) \ \mathbb{E}[|s|^r] \leq r2^{r/2}\sigma_s^r\Gamma(r/2)$$

Let $\mu_v = \mathbb{E}[v]$. By power series expansion and since $\Gamma(r) = (r-1)!$ for an integer $r$, we have:

$$\begin{aligned} \mathbb{E}[e^{t(v-\mu_v)}] &= 1 + t\mathbb{E}[v-\mu_v] + \sum_{r=2}^{\infty} \frac{t^r\mathbb{E}[(v-\mu_v)^r]}{r!} \\ &\leq 1 + \sum_{r=2}^{\infty} \frac{t^r\mathbb{E}[|s|^{2r}]}{r!} \\ &\leq 1 + \sum_{r=2}^{\infty} \frac{t^r 2r2^r\sigma_s^{2r}\Gamma(r)}{r!} \\ &= 1 + \sum_{r=2}^{\infty} t^r 2^{r+1}\sigma_s^{2r} \\ &= 1 + \frac{8t^2\sigma_s^4}{1-2t\sigma_s^2} \end{aligned}$$

By making $|t| \leq 1/(4\sigma_s^2)$, we have $1/(1-2t\sigma_s^2) \leq 2$. Finally, since $(\forall\alpha)\ 1+\alpha \leq e^{\alpha}$, we have that for a sub-Gaussian variable $s$ with parameter $\sigma_s$:

$$(\forall |t| \leq 1/(4\sigma_s^2)) \ \mathbb{E}[e^{t(s^2-\mathbb{E}[s^2])}] \leq e^{16t^2\sigma_s^4} \tag{37}$$

Thus, we obtained a bound for the moment generating function of the sub-exponential variable $s^2$, that is similar to that of sub-Gaussian variables but holds only for a small range of $t$.