Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Risk Minimization in the Presence of Label Noise*

Wei Gao and **Lu Wang** and **Yu-Feng Li** and **Zhi-Hua Zhou**
National Key Laboratory for Novel Software Technology, Nanjing University
Collaborative Innovation Center of Novel Software Technology and Industrialization
Nanjing 210023, China
{gaow, wangl, liyf, zhouzh}@lamda.nju.edu.cn

Abstract

Matrix concentration inequalities have attracted much attention in diverse applications such as linear algebra, statistical estimation, combinatorial optimization, etc. In this paper, we present new Bernstein concentration inequalities depending only on the first moments of random matrices, whereas previous Bernstein inequalities are heavily relevant to the first and second moments. Based on those results, we analyze the empirical risk minimization in the presence of label noise. We find that many popular losses used in risk minimization can be decomposed into two parts, where the first part won't be affected and only the second part will be affected by noisy labels. We show that the influence of noisy labels on the second part can be reduced by our proposed *LICS* (*Labeled Instance Centroid Smoothing*) approach. The effectiveness of the LICS algorithm is justified both theoretically and empirically.

Introduction

Matrix concentration inequalities measure the spectral-norm deviation of a random matrix to its expected mean, and relevant researches have attracted much attention in diverse applications such as statistical estimation (Koltchinskii 2011), linear algebra (Tropp 2011), combinatorial optimization (So 2011), matrix completion (Recht 2011) etc. Various techniques have been developed to study the sum of independent random matrices and matrix martingales (Tropp 2011; 2012; Hsu, Kakade, and Zhang 2012; Mackey et al. 2014). Tropp (2015) made a comprehensive introduction on matrix concentration inequalities.

Empirical risk minimization (Vapnik 1998) has been a popular methodology in diverse learning tasks such as regression, classification, density estimation, etc. In many real applications, the training data often contain noisy labels, e.g., a document may be mis-classified manually due to human error or bias, a doctor may make incorrect diagnoses for patients because of his knowledge and experience, a spanner can manipulate the data to mislead the outcome of span-filter systems, etc. Generally speaking, an empirical risk minimization procedure may be misled by noisy labels. For example, the random noise (Long and Servedio 2010) defeats all convex potential boosters, and support vector machines (SVMs) tend to overfit for noisy labels. It is important to develop effective approaches to make sure that the learning procedure is not misled by noisy data.

In this paper, we first present new matrix Bernstein concentration inequalities depending only on the first moments of random matrices, while previous Bernstein inequalities are heavily relevant to the first and second moments. We further present dimension-free concentration inequalities, which can be used for infinite-dimension matrices. Our new concentration inequalities show tighter bounds for small spectral norm on the first moments of random matrices.

As an application, we utilize new matrix Bernstein concentration inequalities to study the risk minimization of binary classification in the presence of random label noise (also called *random classification noise*). Specifically, the training labels have been flipped with some certain probability instead of seeing true labels. We consider the empirical risk minimization of decomposable losses such as least square loss, logistic loss, etc. The advantage of using such losses is that we can divide empirical risks into two parts, where the first part won't be affected and only the second part will be affected by noisy labels. Further, the risk minimization in the presence of label noise can be converted to the estimation of the statistics *labeled instance centroid*. We prove that label noise can increase the covariance of labeled instance centroid, or even cause heavy-tailed distribution, which makes noisy tasks difficult to learn. We propose the Labeled Instance Centroid Smoothing (LICS) approach to reduce the influence of noisy labels through incorporating labeled instance centroid and its covariance. The effectiveness of LICS is justified both theoretically and empirically.

Related Work

Ahlswede and Winter (2002) possibly proved the first Chernoff concentration inequalities for matrix trace, and similar techniques has been adapted for Bernstein concentration inequalities (Oliveira 2010; Gross 2011). Tropp (2011; 2012; 2015) made fundamental concentration inequalities for random matrices due to (Lieb 1973, Theorem 6). Hsu, Kakade, and Zhang (2012) presented dimension-free con-

*Supported by the 973 Programm (2014CB340501), NSFC (61333014, 61503179) and JiangsuSF (BK20150586).

centration inequalities, where the explicit matrix dimension is replaced by trace quantity. Mackey et al. (2014) derived new exponential concentration inequalities based on the scalar concentration (Chatterjee 2007) via Stein's method of exchangeable pairs.

Angluin and Laird (1988) first proposed the random noise model, and the sample complexity of noise-tolerant learning was studied in (Cesa-Bianchi et al. 1999). Ben-David, Pál, and Shalev-Shwartz (2009) proved that the Little-Stone dimension characterizes the learnability of the online noise learning model. Kearns (1993) introduced the famous statistical query (SQ) model, and Bshouty et al. (1998) presented a SQ algorithm to learn a geometric class in noise-tolerant and distribution-free classification.

Many online approaches have been developed to deal with noise labels, e.g., linear threshold learning (Bylander 1994), passive-aggressive perceptrons (Crammer et al. 2006), confidence-weighted learning (Dredze, Crammer, and Pereira 2008), AROW (Crammer, Kulesza, and Mark 2009), etc. Various non-convex risk minimizations (Xu, Crammer, and Schuurmans 2006; Masnadi-Shirazi and Vasconcelos 2009; 2009; Freund 2009; Denchev et al. 2012) have been developed for noisy labels, and more relevant work can be found in (Frénay and Verleysen 2014). Most of them, however, are heuristic without theoretical guarantees. Manwani and Sastry (2013) made theoretical analysis on the noise-tolerant property of risk minimization of $0/1$ loss and least square loss. Natarajan et al. (2013) suggested unbiased losses for empirical risk minimization, whereas those studies do not consider the influence of variance.

Preliminaries

Let $\mathcal{X}$ and $\mathcal{Y} = \{+1, -1\}$ denote the input and output space, respectively. Let $\mathcal{D}$ be an unknown (noise-free) distribution over $\mathcal{X} \times \mathcal{Y}$. Assume that the training data $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ are drawn identically and independently (i.i.d) according to distribution $\mathcal{D}$.

In the random noise model, each true label y_i is corrupted independently by random noise with rate η, and we denote $\tilde{y}_i$ the corrupted label,

$$\tilde{y}_i = \begin{cases} y_i & \text{with probability } 1-\eta \\ -y_i & \text{with probability } \eta \end{cases}$$

Here η is assumed to be a prior, and it can be estimated via cross-validation in experiments (Natarajan et al. 2013). We focus on uniform noise

$$\Pr[\tilde{y}_i = -1 | y_i = +1] = \Pr[\tilde{y}_i = +1 | y_i = -1] = \eta,$$

and it can be easily generalized to the non-uniform case.

Let $\mathcal{D}_\eta$ be the corrupted distribution, and denote $\tilde{S}_n = \{(\mathbf{x}_1, \tilde{y}_1), (\mathbf{x}_2, \tilde{y}_2), \ldots, (\mathbf{x}_n, \tilde{y}_n)\}$ the corrupted sample by random noise. Let $\mathcal{H} = \{h\colon \mathcal{X} \to \mathbb{R}\}$ be a real-valued function space. For each $h \in \mathcal{H}$, we define the expected risk w.r.t. loss ℓ and true distribution $\mathcal{D}$ as

$$R(h, \mathcal{D}) = E_{(\mathbf{x},y)\sim\mathcal{D}}[\ell(h(\mathbf{x}), y)],$$

where ℓ is a loss function such as least square loss, logistic loss, hinge loss, etc. Further, we define the empirical loss as

$$\hat{R}(h, S_n) = \frac{1}{n}\sum_{i=1}^{n} \ell(h(\mathbf{x}_i), y_i)$$

for $h \in \mathcal{H}$ and $S_n = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$.

Finally, we introduce some notations used in this paper. Let symbol $^\top$ denote the transpose operation on vectors and matrices. For a symmetric matrix X, let $\lambda_{\text{max}}(X)$ and $\lambda_{\text{min}}(X)$ be the largest and smallest eigenvalue, respectively, and $\|X\|$ denotes the spectral norm. For two matrices X_1 and X_2, $X_1 \preceq X_2$ implies that $X_2 - X_1$ is positive semi-definite. For a real number r, let $\lceil r \rceil$ be the smallest integer which is larger than r, and we set $[n] = \{1, \ldots, n\}$ for an integer $n \geq 0$.

New Concentration Inequalities

We begin with new concentration inequalities for random matrix as follows:

Theorem 1 *Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ be i.i.d. random vectors s.t. $\|\mathbf{x}_i\|^2 \leq B$, and write $X_i = \mathbf{x}_i^\top \mathbf{x}_i$. For any $t > 0$, we have*

$$\Pr\left[\left\|\sum_{i=1}^{n} X_i - E[X_i]\right\| \geq t\right] \leq 2de^{-\frac{t^2}{2Bn\lambda_{max}(E[X_1])+Bt}}.$$

This theorem gives new Bernstein concentration inequalities depending only on the first moments of random matrices, whereas previous Bernstein concentration inequalities (Gittens and Tropp 2011; Tropp 2012) are heavily relevant to the first and second moments. This theorem shows tighter bounds for small $\lambda_{\text{max}}(E[X_1])$, i.e., small spectral norm on the first moments of random matrices.

Proof: This proof uses the properties of matrices $\mathbf{x}_i^\top \mathbf{x}_i$ and techniques in (Tropp 2012). For $\theta > 0$ and $i \in [n]$, we have

$$\begin{aligned} E[e^{\theta X_i}] &= I_d + \theta E[X_i] + \sum_{k=2}^{\infty} \frac{\theta^k E[X_i^k]}{k!} \\ E[X_i^k] &= E\left[\|\mathbf{x}_i\|_2^{k-1} X_i\right] \preceq B^{k-1} E[X_i] \end{aligned}$$

for $\|\mathbf{x}_i\|^2 \leq B$ and $\lambda_{\text{max}}(X_i) \geq 0$. If $\theta < 2/B$, we have

$$\begin{aligned} \sum_{k=2}^{\infty} \frac{\theta^k E[X_i^k]}{k!} &\preceq \sum_{k=2}^{\infty} \frac{\theta^k B^{k-1}}{k!} E[X_i] \\ &\preceq \theta \sum_{k=1}^{\infty} (\theta B/2)^k E[X_i] = \frac{\theta^2 B}{2-\theta B} E[X_i] \end{aligned}$$

and

$$E[e^{\theta X_i}] \preceq I_d + \frac{\theta^2 B}{2-\theta B} E[X_i] \preceq e^{\theta E[X_i] + \frac{\theta^2 B}{2-\theta B} E[X_i]}.$$

This yields that $E[e^{\theta(X_i - E[X_i])}] \preceq e^{\frac{\theta^2 B}{2-\theta B} E[X_i]}$ and

$$\begin{aligned} &\Pr\left[\lambda_{\max}\left(\sum_{i=1}^n X_i - E[X_i]\right) \geq t\right] \\ &\leq \inf_{\theta>0} e^{-\theta t} \mathrm{tr}\exp\left(\sum_{i=1}^n \log E[e^{\theta(X_i-E[X_i])}]\right) \\ &\leq \inf_{\theta>0} e^{-\theta t}\mathrm{tr}\exp\left(\tfrac{n\theta^2 B}{2-\theta B}E[X_1]\right) \\ &\leq d\inf_{\theta>0}\exp\left(-\theta t + \tfrac{n\theta^2 B}{2-\theta B}\lambda_{\max}(E[X_1])\right). \end{aligned}$$

By selecting $\theta = 2t/(2Bn\lambda_{\max}(E[X_1]) + Bt)$, it holds that

$$\Pr\left[\lambda_{\max}\Big(\sum_{i=1}^n X_i - E[X_i]\Big) \geq t\right] \leq de^{-\frac{t^2}{2Bn\lambda_{\max}(E[X_1])+Bt}}.$$

We bound $\Pr[\lambda_{\min}(\sum_{i=1}^n X_i - E[X_i]) \leq -t]$ similarly. ■

Theorem 1 presents matrix concentration inequalities, which are explicitly dependent on the matrix dimension d; therefore, those bounds may be looser for high-dimensional matrix. We now present new dimension-free concentration inequalities as follows:

Theorem 2 *Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ denote i.i.d random vectors such that $\|\mathbf{x}_i\|^2 \leq B$. For any $t > 0$, denote by $X_i = \mathbf{x}_i^\top \mathbf{x}_i$, $\gamma = E[\lambda_{max}(X_1)]$, $\alpha = \lambda_{min}(E[X_1])$ and $n_0 = \max(1, \lceil 2t/(\gamma-\alpha)\rceil)$. For $n \geq n_0$, we have*

$$\Pr\left[\left\|\sum_{i=1}^n X_i - E[X_i]\right\| \geq t\right] \leq 2e^{-\frac{t^2}{Bn(\alpha+\gamma)+Bt}}.$$

Proof: By Markov's inequality and for any $\theta > 0$, we have

$$\begin{aligned} &\Pr\left[\lambda_{\max}\Big(\sum_i X_i - E[X_i]\Big) \geq t\right] \\ &\qquad \leq e^{-\theta t} E\left[e^{\theta\lambda_{\max}\left(\sum_i X_i - E[X_i]\right)}\right]. \end{aligned} \quad (1)$$

From $\lambda_{max}(A_1 + A_2) \leq \lambda_{max}(A_1) + \lambda_{max}(A_2)$, we have

$$\begin{aligned} &\lambda_{\max}\Big(\sum_i (X_i - E[X_i])\Big) \leq \sum_i \lambda_{\max}(X_i) \\ &\quad + \lambda_{\max}\Big(-\sum_i E[X_i]\Big) \leq \sum_i \lambda_{\max}(X_i) - n\alpha \end{aligned}$$

where the last inequality holds from $\lambda_{\max}(-\sum_i E[X_i]) = -\lambda_{\min}(\sum_i E[X_i]) = -n\alpha$. Based on Taylor's expansion, it holds that

$$\begin{aligned} e^{\theta\lambda_{\max}(X_i)} &= 1 + \theta\lambda_{\max}(X_i) + \sum_{k\geq 2}\frac{\theta^k \lambda^k_{\max}(X_i)}{k!} \\ &\leq 1 + \theta\gamma + \sum_{k\geq 2}\frac{\theta^k B^{k-1}\gamma}{k!} \end{aligned}$$

where the inequality holds from $\lambda^k_{\max}(X_i) = \lambda_{\max}(X_i^k) = \lambda_{\max}(\|\mathbf{x}_i\|^{2(k-1)}X_i) \leq B^{k-1}\gamma$. For $B\theta \leq 2$, we have

$$1 + \sum_{k\geq 2}\frac{(B\theta)^{k-1}}{k!} \leq 1 + \sum_{k\geq 1}\left(\frac{B\theta}{2}\right)^l = \frac{1}{1-B\theta/2}.$$

This follows that $E[e^{\theta\lambda_{\max}(X_i)}] \leq e^{\theta\gamma/(1-B\theta/2)}$, and

$$e^{-\theta t}E\left[e^{\theta\lambda_{\max}\left(\sum_i (X_i - E[X_i])\right)}\right] \leq e^{-\theta t - \theta n\alpha + \frac{\theta n\gamma}{1-B\theta/2}}. \quad (2)$$

By Eqs. 1 and 2 and setting $\theta = 2(1 - \sqrt{n\gamma/(t+n\alpha)})/B$, it holds that, for $n \geq n_0$,

$$\begin{aligned} &\Pr\left[\lambda_{\max}\Big(\sum_i X_i - E[X_i]\Big) \geq t\right] \\ &\qquad \leq e^{-\frac{2}{B}(\sqrt{t+n\alpha}-\sqrt{n\gamma})^2} \leq e^{-\frac{t^2}{Bn(\alpha+\gamma)+Bt}}. \end{aligned}$$

We bound $\Pr[\lambda_{\min}(\sum_i (X_i - E[X_i])) \leq -t]$ similarly. ■

Theorem 2 gives dimension-free concentration inequalities for random matrix, and can be used for large and infinite-dimension matrix. Hsu, Kakade, and Zhang (2012) presented dimension-free Bernstein concentration inequalities for covariance matrices based on the first and second moments of random matrices. In contrast, our bounds depend only on the first moments of random matrix, and are tighter for small $E[\lambda_{\max}(X_1)]$ and $\lambda_{\min}(E[X_1])$.

Analysis of Label Noise

This section analyzes the risk minimization of decomposable loss, which is defined as follows:

Definition 1 *A loss function ℓ is said to be* decomposable *if there exist some function $g\colon \mathbb{R} \to \mathbb{R}$ and constant c such that the following holds for each $h \in \mathcal{H}$ and S_n*

$$\hat{R}(h, S_n) = \frac{1}{n}\sum_{i=1}^n g(h(\mathbf{x}_i)) + \frac{c}{n}\sum_{i=1}^n y_i h(\mathbf{x}_i).$$

It is easy to show that many loss functions, such as logistic loss and least square loss, are decomposable, and Patrini et al. (Patrini et al. 2014) made similar decomposition for label-proportion learning.

The main advantage of using decomposable losses is that we can divide the empirical loss into two parts, where the first part is not affected but the second part is affected by noisy labels; therefore, it is sufficient to analyze and estimate the influence of second part by noisy labels. For linear classifier $h_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x}\rangle$ and decomposable loss, we have

$$\hat{R}(h_{\mathbf{w}}, S_n) = \frac{1}{n}\sum_{i=1}^n g(\langle \mathbf{x}_i, \mathbf{w}\rangle) + c\Big\langle \frac{1}{n}\sum_{i=1}^n y_i\mathbf{x}_i, \mathbf{w}\Big\rangle.$$

We introduce a new statistics *labeled instance centroid*, with respect to the true sample S_n and true distribution $\mathcal{D}$, as

$$\mu(S_n) = \sum_{i=1}^n \frac{y_i \mathbf{x}_i}{n} \ \text{ and } \ \mu(\mathcal{D}) = E_{(\mathbf{x},y)\sim\mathcal{D}}[y\mathbf{x}].$$

We further define the labeled instance centroid $\mu(\tilde{S}_n)$ and $\mu(\mathcal{D}_\eta)$ with respect to the corrupted sample $\tilde{S}_n$ and corrupted distribution $\mathcal{D}_\eta$ respectively. This follows

$$\hat{R}(h_{\mathbf{w}}, S_n) = c\,\langle \mu(S_n), \mathbf{w}\rangle + \frac{1}{n}\sum_{i=1}^n g(\langle \mathbf{x}_i, \mathbf{w}\rangle).$$

In the random noise model, the true sample S_n, true distribution $\mathcal{D}$ and corrupted distribution $\mathcal{D}_\eta$ are unknown, and what we can observe is a corrupted sample $\tilde{S}_n$. Therefore, the problem of random noise classification can be converted to the estimation of $\mu(S_n)$ from the corrupted sample $\tilde{S}_n$. We present a proposition as follows:

Proposition 1 *We have $\mu(\mathcal{D}_\eta) = (1-2\eta)\mu(\mathcal{D})$ for the true distribution $\mathcal{D}$ and corrupted distribution $\mathcal{D}_\eta$. We have $E_{\tilde{y}_1,\dots,\tilde{y}_n}[\mu(\tilde{S}_n)] = (1-2\eta)\mu(S_n)$ for the true sample S_n and corrupted sample $\tilde{S}_n$.*

Proof: From $E_{\tilde{y}}[\tilde{y}\mathbf{x}|(\mathbf{x},y)] = (1-2\eta)y\mathbf{x}$, we have

$$\begin{aligned}\mu(\mathcal{D}_\eta) &= E_{(\mathbf{x},\tilde{y})\sim\mathcal{D}_\eta}[\tilde{y}\mathbf{x}] = E_{(\mathbf{x},y)\sim\mathcal{D}}[E_{\tilde{y}}[\tilde{y}\mathbf{x}|(\mathbf{x},y)]] \\ &= E_{(\mathbf{x},y)\sim\mathcal{D}}[(1-2\eta)y\mathbf{x}] = (1-2\eta)\mu(\mathcal{D}),\end{aligned}$$

and $E[\mu(\tilde{S}_n)] = (1-2\eta)\mu(S_n)$. ∎

We can see that random noise changes labeled instance centroid, and $\mu(\tilde{S}_n)/(1-2\eta)$ is an unbiased estimation to $\mu(S_n)$. Let $\Sigma(\mathcal{D})$ denote the covariance matrix of the random vector $y\mathbf{x}$ drawn i.i.d. from distribution $\mathcal{D}$, i.e.,

$$\Sigma(\mathcal{D}) = E_{(\mathbf{x},y)\sim\mathcal{D}}\left[[y\mathbf{x}]^\top y\mathbf{x}\right] - [\mu(\mathcal{D})]^\top\mu(\mathcal{D}),$$

and define $\Sigma(\mathcal{D}_\eta)$ similarly. We have

Proposition 2 *We have*

$$\Sigma(\mathcal{D}_\eta) = \Sigma(\mathcal{D}) + 4\eta(1-\eta)[\mu(\mathcal{D})]^\top\mu(\mathcal{D})$$

for the true distribution $\mathcal{D}$ and corrupted distribution $\mathcal{D}_\eta$.

Proof: We have

$$\begin{aligned}\Sigma(\mathcal{D}_\eta) &= E_{(\mathbf{x},\tilde{y})\sim\mathcal{D}_\eta}[(\tilde{y}\mathbf{x})^\top\tilde{y}\mathbf{x}] - [\mu(\mathcal{D}_\eta)]^\top\mu(\mathcal{D}_\eta) \\ &= E_{(\mathbf{x},\tilde{y})\sim\mathcal{D}_\eta}[\mathbf{x}^\top\mathbf{x}] - (1-2\eta)^2[\mu(\mathcal{D})]^\top\mu(\mathcal{D}) \\ &= \Sigma(\mathcal{D}) + 4\eta(1-\eta)[\mu(\mathcal{D})]^\top\mu(\mathcal{D}),\end{aligned}$$

which completes the proof. ∎

This proposition shows that random noise increases the covariance of $y\mathbf{x}$, and may lead to heavy-tailed distributions. For labeled instance centroid $\mu(\tilde{S}_n)$, we consider its covariance matrix $\Sigma(\mu(\tilde{S}_n))$, i.e.,

$$\Sigma(\mu(\tilde{S}_n)) = E[[\mu(\tilde{S}_n)]^\top\mu(\tilde{S}_n)] - [E[\mu(\tilde{S}_n)]]^\top E[\mu(\tilde{S}_n)].$$

We have the following proposition:

Proposition 3 *The covariance matrix $\Sigma(\mu(\tilde{S}_n))$ equals to*

$$E\Big[\sum_{i=1}^n \frac{\mathbf{x}_i^\top\mathbf{x}_i}{n^2}\Big] - E\Big[\sum_{i=1}^n \frac{\mathbf{x}_i^\top\tilde{y}_i}{n^2}\Big]E\Big[\sum_{i=1}^n \frac{\mathbf{x}_i\tilde{y}_i}{n}\Big].$$

Proof: We first have

$$\begin{aligned}E\left[[\mu(\tilde{S}_n)]^\top\mu(\tilde{S}_n)\right] &= E\Big[\Big[\frac{1}{n}\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i\Big]^\top\frac{1}{n}\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i\Big] \\ &= \frac{1}{n^2}\Big(\sum_{i=1}^n E[\mathbf{x}_i^\top\mathbf{x}_i] + \sum_{i\neq j}E[\tilde{y}_i\tilde{y}_j\mathbf{x}_i^\top\mathbf{x}_j]\Big).\end{aligned}$$

For i.i.d random variables $\mathbf{x}_1\tilde{y}_1, \mathbf{x}_2\tilde{y}_2, \dots, \mathbf{x}_n\tilde{y}_n$,

$$E[\tilde{y}_i\tilde{y}_j\mathbf{x}_i^\top\mathbf{x}_j] = E\Big[\sum_{i=1}^n\frac{\mathbf{x}_i\tilde{y}_i}{n}\Big]^\top E\Big[\sum_{i=1}^n\frac{\mathbf{x}_i\tilde{y}_i}{n}\Big]$$

which completes the proof by simple calculation. ∎

Given a corrupted sample $\tilde{S}_n = \{(\mathbf{x}_1,\tilde{y}_1),\dots,(\mathbf{x}_n,\tilde{y}_n)\}$, we define the empirical covariance matrix as

$$\hat{\Sigma}(\mu(\tilde{S}_n)) = \sum_{i=1}^n\frac{\mathbf{x}_i^\top\mathbf{x}_i}{n^2} - \frac{1}{n}\sum_{i=1}^n\frac{\mathbf{x}_i^\top\tilde{y}_i}{n}\sum_{i=1}^n\frac{\mathbf{x}_i\tilde{y}_i}{n}. \quad (3)$$

The following theorem shows that the empirical covariance matrix $\hat{\Sigma}(\mu(\tilde{S}_n))$ is a good approximation of the covariance matrix $\Sigma(\mu(\tilde{S}_n))$.

Theorem 3 *For sample $\tilde{S}_n$, let $\Sigma(\mu(\tilde{S}_n))$ and $\hat{\Sigma}(\mu(\tilde{S}_n))$ be given by Proposition 3 and Eq. 3, respectively. Denote $\gamma = E[\lambda_{max}(\mathbf{x}_1^\top\mathbf{x}_1)]$, $\alpha = \lambda_{min}(E[\mathbf{x}_1^\top\mathbf{x}_1])$, $\tau = tr(E[\mathbf{x}_1^\top\mathbf{x}_1])$. For $t>0$, we set $n_0 = \max(1, \lceil 2t/(\gamma-\alpha)\rceil)$. For $n \geq n_0$, it holds that, with probability at least $1-3e^{-t}$*

$$\begin{aligned}&\left\|\Sigma(\mu(\tilde{S}_n)) - \hat{\Sigma}(\mu(\tilde{S}_n))\right\| \\ &\qquad\leq \frac{11Bt}{3n^2} + \frac{\sqrt{Bt(\alpha+\gamma)} + \sqrt{2B\tau}(1+\sqrt{8t})}{n^{3/2}}.\end{aligned}$$

Proof: Based on Proposition 3 and Eq. 3, we first give the upper bound for $\|\Sigma(\mu(\tilde{S}_n)) - \hat{\Sigma}(\mu(\tilde{S}_n))\|$ as follows:

$$\begin{aligned}&\frac{1}{n^3}\Big\|E\Big[\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i^\top\Big]E\Big[\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i\Big] - \sum_{i=1}^n\tilde{y}_i\mathbf{x}_i^\top\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i\Big\| \\ &\qquad + \frac{1}{n^2}\Big\|\sum_{i=1}^n(E[\mathbf{x}_i^\top\mathbf{x}_i] - \mathbf{x}_i^\top\mathbf{x}_i)\Big\|\end{aligned}$$

For $t>0$ and $n\geq n_0$, Theorem 2 shows that

$$\frac{1}{n^2}\Big\|\sum_{i=1}^n(E[\mathbf{x}_i^\top\mathbf{x}_i] - \mathbf{x}_i^\top\mathbf{x}_i)\Big\| \leq \frac{Bt+\sqrt{Bnt(\alpha+\gamma)}}{n^2}$$

with probability at least $1-2e^{-t}$. For $\|\mathbf{x}_i\|^2 \leq B$, we have

$$\begin{aligned}&\Big\|E\Big[\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i^\top\Big]E\Big[\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i\Big] - \Big[\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i^\top\Big]\Big[\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i\Big]\Big\| \\ &\qquad\leq \sqrt{2B}n\Big\|\sum_{i=1}^n\tilde{y}_i\mathbf{x}_i - E[\tilde{y}_i\mathbf{x}_i]\Big\|.\end{aligned}$$

This follows $E[\tilde{y}_i\mathbf{x}_i - E[\tilde{y}_i\mathbf{x}_i]] = 0$, $\|\tilde{y}_i\mathbf{x}_i - E[\tilde{y}_i\mathbf{x}_i]\| \leq \sqrt{2B}$ and $E[\|\tilde{y}_i\mathbf{x}_i - E[\tilde{y}_i\mathbf{x}_i]\|^2] \leq \tau$. By Bernstein bounds (Hsu, Kakade, and Zhang 2012), we have, with probability at least $1-e^{-t}$

$$\Big\|\sum_{i=1}^n(\tilde{y}_i\mathbf{x}_i - E[\tilde{y}_i\mathbf{x}_i])\Big\| \leq \sqrt{n\tau}(1+\sqrt{8t}) + \frac{4\sqrt{2B}}{3}t. \quad (4)$$

This completes the proof by simple calculations. ∎

Algorithm 1 Median-of-means estimator of label mean

Input: The corrupted sample $\tilde{S}_n$, number of groups $k \geq 1$.
Output: Median-of-means estimator $\hat{\mu}(\tilde{S}_n)$.

1: Randomly partition $\tilde{S}_n$ into k groups $\tilde{S}_n^{[1]}, \tilde{S}_n^{[2]}, \ldots, \tilde{S}_n^{[k]}$ such that each group has almost equal size.
2: Calculate the standard empirical mean $\mu(\tilde{S}_n^{[i]})$ for each $i \in [k]$ and each group $\tilde{S}_n^{[i]}$.
3: Calculate $r_i = \text{median}_j\{\|\mu(\tilde{S}_n^{[i]}) - \mu(\tilde{S}_n^{[j]})\|\}$ for each $i \in [k]$, and then set $i_* = \arg\min_{i\in[k]} r_i$.
4: Return $\hat{\mu}(\tilde{S}_n) = \mu(\tilde{S}_n^{[i_*]})$.

The LICS Algorithm

Proposition 2 shows that random noise increases the covariance of $y\mathbf{x}$, and may lead to heavy-tailed distributions. We adopt the recent generalized median-of-means estimator (Hsu and Sabato 2014), rather than using the standard empirical mean, to estimate the corrupted labeled instance centroid $\mu(\tilde{S}_n)$. The basic idea is to randomly partition the corrupted sample $\tilde{S}_n$ into k groups with almost equal size, and return the generalized median of sample means for each group under L_2-norm metric. The detailed description is presented in Algorithm 1.

We further consider a range $\mathcal{R}$ for $\mu(\tilde{S}_n)$ as follows:

$$\mathcal{R} = \{\mu\colon (\mu - \hat{\mu}(\tilde{S}_n))^\top \hat{\Sigma}(\mu(\tilde{S}_n))(\mu - \hat{\mu}(\tilde{S}_n)) \leq \beta\} \quad (5)$$

where $\hat{\mu}(\tilde{S})$ is the output of Algorithm 1, $\hat{\Sigma}(\mu(\tilde{S}))$ is defined by Eq. 3, and β is a parameter estimated by cross-validation. Our optimization problem can be formalized as

$$\min_{\mathbf{w},\mu} \ \frac{1}{n}\sum_{i=1}^{n} g(\langle \mathbf{x}_i, \mathbf{w}\rangle) + \frac{c\,\langle \mathbf{w}, \mu\rangle}{1-2\eta} + \lambda\|\mathbf{w}\|^2 \quad (6)$$
$$\text{s.t. } (\mu - \hat{\mu}(\tilde{S}_n))^\top \hat{\Sigma}(\mu(\tilde{S}_n))(\mu - \hat{\mu}(\tilde{S}_n)) \leq \beta.$$

We will employ an alternating method to address such optimization. Specifically, when μ is fixed, we need to solve

$$\min_{\mathbf{w}} \ \frac{1}{n}\sum_{i=1}^{n} g(\langle \mathbf{x}_i, \mathbf{w}\rangle) + \frac{c\,\langle \mathbf{w}, \mu\rangle}{1-2\eta} + \lambda\|\mathbf{w}\|^2.$$

This minimization can be optimized by standard and simple gradient descent algorithm. For fixed $\mathbf{w}$, it sufficient to solve

$$\min_{\mu} \quad c\,\langle \mathbf{w}, \mu\rangle \quad (7)$$
$$\text{s.t. } (\mu - \hat{\mu}(\tilde{S}_n))^\top \hat{\Sigma}(\mu(\tilde{S}_n))(\mu - \hat{\mu}(\tilde{S}_n)) \leq \beta,$$

and we can give a closed-form solution for this problem. By introducing a Lagrange variable ρ, we have

$$L(\mu, \beta) = c\,\langle \mathbf{w}, \mu\rangle - \rho(\mu - \hat{\mu}(\tilde{S}_n))^\top \hat{\Sigma}(\mu(\tilde{S}_n))(\mu - \hat{\mu}(\tilde{S}_n)) + \rho\beta.$$

By solving $\partial L(\mu, \beta)/\partial\mu = 0$, we have

$$\mu = \frac{c}{2\rho}\big(\hat{\Sigma}(\mu(\tilde{S}_n))\big)^{-1}\mathbf{w} + \hat{\mu}(\tilde{S}_n) \quad (8)$$

Algorithm 2 The Labeled Instance Centroid Smooth (LICS) algorithm

Input: The corrupted sample $\tilde{S}_n = \{(\mathbf{x}_1, \tilde{y}_1), (\mathbf{x}_2, \tilde{y}_2), \ldots, (\mathbf{x}_n, \tilde{y}_n)\}$, the noisy parameter η, the regularization parameter λ, the approximation parameter β.
Output: The classifier $\mathbf{w}_t$.

1: Call Algorithm 1 to give an estimation of $\hat{\mu} = \hat{\mu}(\tilde{S}_n)$.
2: Calculate $\hat{\Sigma} = \hat{\Sigma}(\mu(\tilde{S}_n))$ by Eq. 3.
3: Initialize $t = 1$ and $\mathbf{w}_0$.
4: Calculate $\mu = \hat{\mu} + \hat{\Sigma}^{-1}\mathbf{w}_{t-1}\sqrt{\beta/\mathbf{w}_{t-1}\hat{\Sigma}^{-1}\mathbf{w}_{t-1}^\top}$.
5: Update $t = t + 1$ and solve
$$\mathbf{w}_t = \arg\min_{\mathbf{w}} \sum_{i=1}^{n} \frac{g(\langle \mathbf{x}_i, \mathbf{w}\rangle)}{n} + \frac{c\,\langle \mathbf{w}, \mu\rangle}{1-2\eta} + \lambda\|\mathbf{w}\|^2.$$
6: Repeat Steps 4 and 5 until convergence.
7: Return $\mathbf{w}_t$

Substituting Eq.(8) into Eq.(7) and ignoring some constant terms, we have,

$$\min_{\rho} \quad \frac{c^2}{2\rho}\,\mathbf{w}^\top\big(\hat{\Sigma}(\mu(\tilde{S}_n))\big)^{-1}\mathbf{w}$$
$$\text{s.t.} \quad \frac{c^2}{4\rho^2}\,\mathbf{w}^\top\big(\hat{\Sigma}(\mu(\tilde{S}_n))\big)^{-1}\mathbf{w} \leq \beta.$$

We get $\rho = -\frac{c}{2}(\mathbf{w}^\top\big(\hat{\Sigma}(\mu(\tilde{S}_n))\big)^{-1}\mathbf{w}/\beta)^{1/2}$. Substituting to Eq.(8), we derive the optimal solution of Eq.(7) as

$$\mu = \hat{\mu}(\tilde{S}_n) + \big(\hat{\Sigma}(\mu(\tilde{S}_n))\big)^{-1}\mathbf{w}\sqrt{\beta/\mathbf{w}^\top\big(\hat{\Sigma}(\mu(\tilde{S}_n))\big)^{-1}\mathbf{w}}.$$

Algorithm 2 shows the detail procedures of our algorithm.

Theoretical Guarantee for the LICS Algorithm

For $\|\mathbf{x}\|^2 \leq B$ and $\mathcal{H} = \{\mathbf{w}\colon \|\mathbf{w}\|^2 \leq B_0\}$, we denote by $\mathfrak{R}(\mathcal{H}) = E_{\mathbf{x}_i,\epsilon_i}[\sup_{\mathbf{w}\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i\mathbf{w}^\top\mathbf{x}_i]$ the Rademacher complexity. Kakade, Sridharan, and Tewari (2008) showed that $\mathfrak{R}(\mathcal{H}) \leq \sqrt{BB_0/n}$. We have the following theoretical guarantee for the proposed LICS algorithm:

Theorem 4 *For corrupted sample $\tilde{S}_n$ such that $\|\mathbf{x}_i\|^2 \leq B$, let $(\tilde{\mathbf{w}}, \tilde{\mu})$ be the optimal solution of Eq. 6 and set $\tau = tr(E[\mathbf{x}_1^\top\mathbf{x}_1])$. For $t > 0$, there exists $n_0 > 0$ such that, for $n \geq n_0$, the following holds with probability at least $1 - 6e^{-t}$*

$$R(\tilde{\mathbf{w}}, \mathcal{D}) \leq \min_{\mathbf{w}\in\mathcal{H}} R(\mathbf{w}, \mathcal{D}) + 4L\sqrt{BB_0/n} + \sqrt{2t/n} + \frac{\delta_1}{\sqrt{n}}(1 + \sqrt{8t}) + \frac{\delta_2}{3n}t + \frac{1}{1-2\eta}\sqrt{\frac{\beta}{\lambda_{min}(\hat{\Sigma}(\mu(\tilde{S}_n)))}}$$

where $\delta_1 = |c|(\sqrt{\tau B_0} + \sqrt{k\tau B_0})/(1-2\eta)$, $\delta_2 = 4(1+k)\sqrt{2B}/((1-2\eta))$, k is the group number in Algorithm 1.

The parameter β is set as $O(1/n)$ in experiments so as to guarantee the convergence of proposed algorithm since the covariances of labeled instance centroid converges at $O(1/n)$ rate.

Table 1: Comparison of test accuracies (mean±std.) for various approaches on UCI benchmark datasets. •/∘ indicates that LICS is significantly better/worse than the corresponding method (paired t-tests at 95% significance level).

Dateset (#dim., #inst.)	η	LICS	ULE	AROW	PA-II	NTP
australian (14, 690)	0.1	0.8643±0.0280	0.8632±0.0282	0.8626±0.0307	0.8597±0.0286	0.8400±0.0492•
	0.2	0.8530±0.0301	0.8538±0.0305	0.8507±0.0307	0.8580±0.0374	0.8275±0.0637•
	0.3	0.8480±0.0328	0.8442±0.0386	0.8248±0.0426•	0.8441±0.0484	0.8052±0.0743•
	0.4	0.8062±0.0489	0.7857±0.0660	0.7783±0.0546	0.7852±0.0853	0.7006±0.1322•
breast (10, 683)	0.1	0.9608±0.0134	0.9584±0.0150	0.9579± 0.0142	0.9555±0.0170	0.9491±0.0271
	0.2	0.9557±0.0168	0.9546±0.0193	0.9526± 0.0198	0.9347±0.0238•	0.9470±0.0261
	0.3	0.9473±0.0238	0.9258±0.0252•	0.9218± 0.0242•	0.9330±0.0282	0.9304±0.0437
	0.4	0.9286±0.0310	0.9014±0.0508•	0.9025± 0.0353•	0.8946±0.0551•	0.8097±0.1016•
diabetes (8, 768)	0.1	0.7702±0.0342	0.7696±0.0406	0.7667±0.0381	0.7399±0.0399•	0.7380±0.0424•
	0.2	0.7563±0.0349	0.7418±0.0391	0.7321±0.0384•	0.6982±0.0501•	0.7154±0.0430•
	0.3	0.7492±0.0482	0.7238±0.0534•	0.7213±0.0530•	0.6634±0.0906•	0.6457±0.1211•
	0.4	0.7202±0.0571	0.6809±0.0635•	0.6981 ± 0.0633	0.6286±0.1096•	0.5983±0.1305•
german (24, 1000)	0.1	0.7686±0.0186	0.7506±0.0207•	0.7568±0.0212	0.7356±0.0355•	0.7082±0.0768•
	0.2	0.7499±0.0220	0.7426±0.0260	0.7424±0.0252	0.7280±0.0339•	0.6884±0.0721•
	0.3	0.7280±0.0205	0.7048±0.0346•	0.7024±0.0335•	0.7026±0.0371•	0.6476±0.1366•
	0.4	0.7071±0.0305	0.6918±0.0329	0.6584±0.0434•	0.7002±0.0323	0.5904±0.1475•
heart (13, 270)	0.1	0.8231±0.0332	0.8289±0.0448	0.8193±0.0488	0.8311±0.0442	0.8119±0.0544
	0.2	0.8102±0.0412	0.8148±0.0507	0.7963±0.0431	0.8052±0.0490	0.7815±0.0600
	0.3	0.8007±0.0546	0.7970±0.0567	0.7504±0.0596	0.7919±0.0608	0.7452±0.0880•
	0.4	0.7538±0.0747	0.7178±0.0825•	0.6756±0.0823•	0.7059±0.0980•	0.6378±0.1332•
splice (60, 1000)	0.1	0.7986±0.0295	0.7968±0.0328	0.7970±0.0340	0.7712±0.0336	0.7256±0.0501•
	0.2	0.7597±0.0316	0.7606±0.0331	0.7554±0.0325	0.7560±0.0396	0.6956±0.0656•
	0.3	0.7208±0.0387	0.7003±0.0397•	0.7063±0.0388	0.7096±0.0419	0.6526±0.0572•
	0.4	0.6613±0.0418	0.6398±0.0459•	0.6252±0.0426•	0.6544±0.0518	0.5722±0.0465•

Proof: Suppose that S_n is the corresponding true sample without corruption. Let $\hat{\mathbf{w}} = \arg\min_{\mathbf{w}\in\mathcal{H}} \hat{R}(\mathbf{w}, S_n)$ and $\mathbf{w}^* = \arg\min_{\mathbf{w}\in\mathcal{H}} R(\mathbf{w}, \mathcal{D})$. Bartlett and Mendelson (2002) proved that

$$\begin{aligned} & R(\tilde{\mathbf{w}}, \mathcal{D}) - R(\mathbf{w}^*, \mathcal{D}) \\ \leq \quad & \hat{R}(\tilde{\mathbf{w}}, S_n) - \hat{R}(\hat{\mathbf{w}}, S_n) + 4L\mathfrak{R}(\mathcal{H}) + 2\sqrt{t/2n}. \end{aligned}$$

Let $\hat{\mu}$ denote the output of Algorithm 1 with input $\tilde{S}_n$ and group number k. We have $(1-2\eta)E[\mu(S_n)] = E[\hat{\mu}]$. Write $\widetilde{R}(\mathbf{w}, \mu)$ as the objective function in Eq. 6, i.e.,

$$\widetilde{R}(\mathbf{w}, \mu) = \frac{c\langle \mathcal{H}, \mu\rangle}{1-2\eta} + \frac{1}{n}\sum_{i=1}^{n} g(\langle \mathbf{x}_i, \mathcal{H}\rangle) + \lambda\|\mathcal{H}\|^2,$$

and $(\tilde{\mathbf{w}}, \tilde{\mu}) = \arg\min_{\mathbf{w}\in\mathcal{H}, \mu\in\mathcal{R}} \widetilde{R}(\mathbf{w}, \mu)$. Similarly to the proof of Eq. 4, it holds with probability at least $1 - 2e^{-t}$,

$$\begin{aligned} & \|(1-2\eta)\mu(S_n) - \hat{\mu}\| \\ & \qquad \leq \frac{\sqrt{\tau} + \sqrt{k\tau}}{\sqrt{n}}(1 + \sqrt{8t}) + \frac{4(1+k)\sqrt{2B}}{3n} t. \quad (9) \end{aligned}$$

Thus, there is a $n_0 > 0$ such that $(1-2\eta)\mu(S_n) \in \mathcal{R}$ for $n \geq n_0$, which yields $\widetilde{R}(\tilde{\mathbf{w}}, \tilde{\mu}) \leq \hat{R}_\ell(\hat{\mathcal{H}}, S_n)$. For $\tilde{\mu} \in \mathcal{R}$, we have $\|\hat{\mu} - \tilde{\mu}\| \leq \sqrt{\beta/\lambda_{\min}(\hat{\Sigma}(\mu(\tilde{S}_n)))}$. This completes the proof by combining with Eq. 9. ■

Experiments

We evaluate the performance of the LICS algorithm on six UCI[1] datasets: australian, breast, diabetes, german, heart and splice. Most of them have been investigated in previous work, and all features are scaled to $[-1, 1]$. We compare the proposed LICS algorithm with four state-of-the-art noisy approaches: unbiased logistic estimator (ULE) classifier (Natarajan et al. 2013), AROW (Crammer, Kulesza, and Mark 2009), passive-aggressive II algorithm (PA-II) (Crammer et al. 2006) and noise-tolerant percetron (NTP) (Khardon and Wachman 2007). In the proposed LICS algorithm, five-fold cross-validation is executed to select the regularized parameter $n\lambda \in \{2^{-12}, 2^{-11}, \ldots, 2^2\}$ (n is size of training data), approximation parameter $n\beta \in \{2^{-12}, 2^{-12}, \ldots, 2^{12}\}$, noise rate $\eta \in \{0.1, 0.2, 0.3, 0.4\}$, and we set group number $k = 3$ in Algorithm 1. The parameters in all compared methods are chosen by cross-validation in a similar manner.

The performance is evaluated by five trials of 5-fold cross validation, and the test accuracies are obtained by averaging over these 25 runs, as summarized in Table 1. We can see that the proposed LICS achieves better or comparable performance, as well as smaller variance, over all datasets. One possible reason is that LICS considers a range $\mathcal{R}$ for estimated labeled instance centroid (Eq. 5) and derives a smaller empirical risk for noisy label, rather than simply taking the estimated labeled instance centroid as the ground-truth.

Conclusion

Matrix concentration inequalities have attracted much attention in diverse applications. This paper presents new Bernstein concentration inequalities depending only on the first moments of random matrices, whereas previous Bernstein concentration inequalities are heavily relevant to the first and

[1] http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

second moments. We further analyze the empirical risk minimization in the presence of label noise. We find that many popular losses used in empirical risk minimization can be decomposed into two parts, where the first part won't be affected and only the second part will be affected by noisy labels. We show that the influence of noisy labels on the second part can be reduced by our proposed LICS approach, and the effectiveness of LICS is justified both theoretically and empirically. It is interesting to presents tighter matrix concentration inequalities and extend the LICS approach to other losses such as exponential loss and hinge loss for future researches.

References

Ahlswede, R., and Winter, A. 2002. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory* 48(3):569–579.

Angluin, D., and Laird, P. 1988. Learning from noisy examples. *Machine Learning* 4(2):343–370.

Bartlett, P. L., and Mendelson, S. 2002. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3:463–482.

Ben-David, S.; Pál, D.; and Shalev-Shwartz, S. 2009. Agnostic online learning. In *Proceedings of the 22nd Conference on Learning Theory*.

Bshouty, N.; Goldman, S.; Mathias, H.; Suri, S.; and Tamaki, H. 1998. Noise-tolerant distribution-free learning of general geometric concepts. *Journal of the ACM* 45(5):863–890.

Bylander, T. 1994. Learning linear threshold functions in the presence of classification noise. In *Proceedings of the 7th Annual Conference on Computational Learning Theory*, 340–347.

Cesa-Bianchi, N.; Dichterman, E.; Fischer, P.; Shamir, E.; and Simon, H. 1999. Sample-efficient strategies for learning in the presence of noise. *Journal of the ACM* 46(5):684–719.

Chatterjee, S. 2007. Steins method for concentration inequalities. *Probability theory and related fields* 138(1):305–321.

Crammer, K.; Dekel, O.; Keshet, J.; Shalev-Shwartz, S.; and Singer, Y. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research* 7:551–585.

Crammer, K.; Kulesza, A.; and Mark, D. 2009. Adaptive regularization of weight vectors. In *Advances in Neural Information Processing Systems 22*. Cambridge, MA: MIT Press. 414–422.

Denchev, V.; Ding, N.; Neven, H.; and Vishwanathan, S. 2012. Robust classification with adiabatic quantum optimization. In *Proceedings of the 29th International Conference on Machine Learning*, 863–870.

Dredze, M.; Crammer, K.; and Pereira, F. 2008. Confidence-weighted linear classification. In *Proceedings of the 25th International Conference on Machine Learning*, 264–271.

Frénay, B., and Verleysen, M. 2014. Classification in the presence of label noise: A survey. *IEEE transactions on neural networks and learning systems* 25(5):845–869.

Freund, Y. 2009. A more robust boosting algorithm. *CoRR/abstract* 0905.2138.

Gittens, A., and Tropp, J. A. 2011. Tail bounds for all eigenvalues of a sum of random matrices. *CoRR* abs/1104.4513v2.

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

Hsu, D., and Sabato, S. 2014. Heavy-tailed regression with a generalized median-of-means. In *Proceedings of the 31th International Conference on Machine Learning*, 37–45.

Hsu, D.; Kakade, S. M.; and Zhang, T. 2012. Tail inequalities for sums of random matrices that depend on the intrinsic dimension. *Electronic Communications in Probability* 17(14):1–13.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2008. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Advances in Neural Information Processing Systems 24*. Cambridge, MA: MIT Press. 793–800.

Kearns, M. 1993. Efficient noise-tolerant learning from statistical queries. In *Proceedings of the 25th Annual ACM Symposium on Theory of Computing*, 392–401.

Khardon, R., and Wachman, G. 2007. Noise tolerant variants of the perceptron algorithm. *Journal of Machine Learning Research* 8:227–248.

Koltchinskii, V. 2011. Oracle inequalities in empirical risk minimization and sparse recovery problems. *Lecture Notes in Math. 2033*.

Lieb, E. 1973. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Mathematics* 11(3):267–288.

Long, P., and Servedio, R. 2010. Random classification noise defeats all convex potential boosters. *Machine Learning* 78(3):287–304.

Mackey, L.; Jordan, M.; Chen, R.; Farrell, B.; and Tropp, J. 2014. Matrix concentration inequalities via the method of exchangeable pairs. *Annals of Probability* 42(3):906–945.

Manwani, N., and Sastry, P. 2013. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics* 43(3):1146–1151.

Masnadi-Shirazi, H., and Vasconcelos, N. 2009. On the design of loss functions for classification: Theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 22*. Cambridge, MA: MIT Press. 1049–1056.

Natarajan, N.; Dhillon, I.; Ravikumar, P.; and Tewari, A. 2013. Learning with noisy labels. In *Advances in Neural Information Processing Systems 26*. Cambridge, MA: MIT Press. 1196–1204.

Oliveira, R. 2010. Sums of random hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability* 15:203–212.

Patrini, G.; Nock, R.; Caetano, T.; and Rivera, P. 2014. (almost) no label no cry. In *Advances in Neural Information Processing Systems 27*. Cambridge, MA: MIT Press. 190–198.

Recht, B. 2011. A simpler approach to matrix completion. *Journal of Machine Learning Research* 12:3413–3430.

So, A. 2011. Moment inequalities for sums of random matrices and their applications in optimization. *Mathematical programming* 130(1):125–151.

Tropp, J. 2011. Freedman's inequality for matrix martingales. *Electronic Communications in Probability* 16:262–270.

Tropp, J. 2012. User-Friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics* 12(4):389–434.

Tropp, J. 2015. An introduction to matrix concentration inequalities. *CoRR* abs/1501.01571.

Vapnik, V. N. 1998. *Statistical Learning Theory*. John Wiley & Sons.

Xu, L.; Crammer, K.; and Schuurmans, D. 2006. Robust support vector machine training via convex outlier ablation. In *Proceedings of the 21st Conference on Artificial Intelligence*, 536–542.